UNIVERSITY OF SOUTHAMPTON

Conformational sampling of intrinsically disordered peptides by enhanced sampling methods

by

Marija Miljak

October 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Faculty of Natural and Environmental Sciences Chemistry

Doctor of Philosophy

Conformational sampling of intrinsically disordered peptides by enhanced sampling methods

by Marija Miljak

The aim of this study was to explore the conformational equilibrium of four cyclic hormone peptides in order to investigate to what extent the bound conformational state can be observed from the solution phase simulations. The studied cyclic peptides share the same structural motif of a six membered ring closed by disulphide bridge between the cysteine residues. They also belong to the class of intrinsically disordered peptides known to exist in an equilibrium of different conformations. Elucidating their conformational ensemble using traditional experimental techniques has proven hard due to the fast interconversion between conformational states, and thus molecular dynamics simulation may help in providing a detailed picture of the peptide's conformational ensemble.

However, conventional molecular dynamics simulation are limited by the long time scale required to observe many conformational motions. Therefore in this work Replica Exchange techniques were applied to test the rate of convergence in conformational sampling. Moreover, to predict the conformational equilibrium of the peptides, a combination of results from enhanced sampling methods, DFT calculations and NMR experiments was used. It was found that calculated chemical shifts weighted by the ensemble populations of each conformational state were better able to reproduce the experimental chemical shift data, over and above any single peptide conformation. This result supports the use of enhanced sampling molecular dynamics computer simulations to study intrinsically disordered peptides.

The knowledge of the conformational equilibrium and the relative populations of the unbound states of the peptides obtained using this approach may help in predicting the structural and functional roles of the bound state peptide. Another purpose of this work was also to check the extent to which a difference in peptide sequence may contribute to their functional diversity. Finally, the performance of

the Replica Exchange simulations was compared, indicating that Solute Tempering is to be preferred over temperature Replica Exchange for reasons of computational efficiency.

Acknowledgements

Here I would like to thank to the several people who have been by my side during this journey. First of all, I would like to thank my supervisor, Jon Essex for giving me the opportunity to come to Southampton to do the PhD. Thank you for all the support, guidance and patience during these four years. I would also like to thank our collaborators from the University of Portsmouth for always being very enthusiastic about this project, and sharing their knowledge with us. In particular, I would like to thank to Elke Haensele, whose feedback was very valuable at the different stages of this project, and who was always ready to help me at any time.

A special thank you goes to my sweet friend Nawel Mele. I am so happy our lives crossed during this PhD. I will never forget your help from the first days in Southampton to all the ups and downs we have been through together in the mean time. Having you by my side, made this journey very special. I have been enjoying every minute of our friendship ever since, thank you for everything!

I would also like to thank my parents for always being very supportive and caring, even though they were not physically close to me. Of course, big thank you to my sister for always keeping her eye on me and putting smile on my face. My thank you also extends to all my friends in Croatia, but also to the friends I made here.

After all, I would like to thank all the members of the Essex group for all the fun, friendship and help during my time in Southampton.

Contents

List of Tables

List of Figures

Chapter 1

Introduction - The structure and biological function of peptides

1.1 The structure of peptides

Peptides are short chains of amino acids linked by peptide bond. They are composed of up to 50 amino acids, in comparison to proteins whose chain length can exceed 100 residues [1]. A peptide bond is defined as a bond between the carbonyl carbon and the amide nitrogen. It is of partial double bond character due to the delocalisation of the lone electron pair from nitrogen, which restrains the peptide bond rotation and makes it planar. This causes peptide bond to be in either *cis* or *trans* configuration [2]. In nature, the *trans* conformation of the peptide bond is favoured over *cis* conformation due to steric effect arising between two alpha carbons being on the same side [3]. The *cis/trans* configuration of the peptide bond is described by ω torsion angle defined between $CA_i - C_i - N_{i+1} - CA_{i+1}$ atoms. A *cis* conformation takes values of ω close to 0 deg, and around ± 180 deg for *trans* configuration of the peptide bond.

A special case is a peptide bond next to proline, usually referred to as Xaa-Pro where Xaa is any amino acid. Due to the tertiary nature of the amide nitrogen in proline, *cis/trans* isomerisation is almost equally energetically favoured [4, 5].

Peptide backbone flexibility is described by extra two angles, ϕ and ψ. The ϕ is describing the rotation around C-N-CA-C bond, and ψ around N-CA-C-N bond (Figure 1.4). The ϕ and ψ torsion angles are usually used to describe secondary structure motifs in peptides. All values of the $\phi\psi$ angles are possible in the −180° to 180° angle range, but due to the steric constraints between atoms in the polypeptide backbone and amino acid side chains, not all values are allowed. The $\phi\psi$ angles distribution is usually given by the Ramachandran plot [6], which provides an easy view of the energetically allowed regions of the $\phi\psi$ angles in proteins (Figure 1.1).

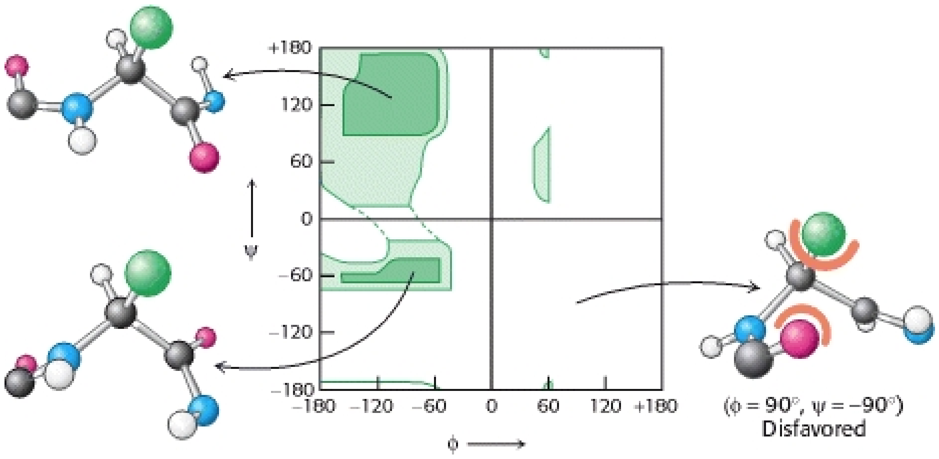


Figure 1.1: A Ramachandran plot showing the allowed regions of $\phi\psi$ angles in green for the given molecule. The structure on the right is disfavoured due to the steric clashes between the atoms surrounded by red semicircles [7].

A sequence of amino acids building a particular protein is known as the *primary* protein structure. Other structural levels of organisation distinguished in proteins include *secondary* structure which describes particular structural segments, *tertiary* structure which defines three dimensional (3D) structure of protein, and *quaternary* which refers to the interactions between domains belonging to the same protein chain, and interactions formed by distinct protein chains [8].

1.2 Secondary structure elements

In the context of this work, secondary structure motifs are particularly relevant. There are several types of secondary structures found in proteins, such as alpha helices, beta sheets, turns and coils [9].

A **coil** is usually referred to as a sequence of amino acids that are neither helix, beta sheet or turn.

An **alpha helix** is secondary structure type often found in larger proteins (Figure 1.2 (b)). It is defined between four consecutive residues forming a 3.6 turn connected with hydrogen bond between amine at residue i and carboxyl group at residue $i+4$ [10]. This ideal helix is also referred to as 3.6_{13} helix, where 3.6 is the number of residues per turn, and 13 is the number of atoms in one turn. Another type of helix occurring in proteins is 3_{10} helix with turn made of three residues, and a hydrogen bond between residues i and $i+3$, instead of $i+4$ for ideal alpha helix.

Another secondary structure type often found in proteins is **beta sheet** (Figure 1.2 (a)), where several protein beta strands are joined edge to edge in the opposite direction (antiparallel) or in the same direction (parallel) to form a sheet where the CO group of each amino acid beta strand is bound by hydrogen bonds with the NH group of the other strand [11].

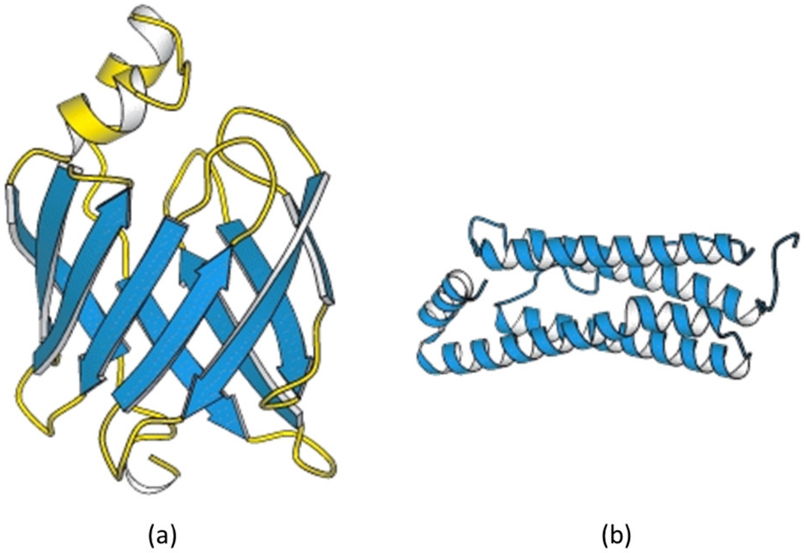


Figure 1.2: (a) A protein rich in beta sheets (blue arrows), (b) An alpha helical protein [7].

While alpha helices and beta sheets are characterised with repetitive motifs stabilised by hydrogen bonds, another type of secondary structure element, called turn, is characterised by a particular range of $\phi\psi$ torsion angles. There are two types of turns, β-turn and γ-turn. More detailed description of the turn types is given in the next section.

1.2.1 β-turns

A β-turn was first recognised as a secondary structure motif by Venkatachalam [12] who was looking for conformational pattern occurring in a system linked by three consecutive peptide bonds that could be stabilised by the hydrogen bonds between residues *i, i+3*.

He discovered three different β-turn types depending on the value of ϕ and ψ torsion angles of the residues *i+1* and *i+2* (I,II,III). Type III is simply 3_{10} helix, already described in the previous section. Types I and III β-turns differ only by 30° for the values of angles $\phi(i+2)$ and $\psi(i+2)$ (Table 1.1).

However, not all β-turns posses a hydrogen bond, so Lewis et al. [13] suggested a new criterion for β-turn definition imposing the requirement that the distance between the C_α (i) and the C_α $(i+3)$ was $<$ 7 Å and the residues involved were not helical. If a hydrogen bond is not present and $\phi\psi$ torsion angles are varying +/- 30° from the ideal turn values, then turn type is referred to as *open* [14].

In addition to types I, II and III, each of these three turns types also has a backbone mirror-image conformation I', II', and III'.

Other turn types include type VI β-turns which differ from other turn types because they involve *cis*-proline peptide bond at *i+2* position, additionally divided into subtypes VIa and VIb [9] (Table 1.1).

The list of the most common β-turn types is given in Table 1.1 [9], while the difference between some of them is visually shown in Figure 1.3.

Type	ϕ (i+1)	ψ (i+1)	ϕ (i+2)	ψ (i+2)
I	-60	-30	-90	0
II	-60	120	80	0
III	-60	-30	-60	-30
I'	60	30	90	0
II'	60	-120	-80	0
VIII	-60	-30	-120	120
VIa	-60	120	-90	0
VIb	-135	135	-75	160

Table 1.1: Type of β-turns and their ideal ϕ and ψ angles. The values of the ϕ and ψ angles are allowed to vary +/- 30° from the ideal turn values.

1.2.2 γ-turn

A γ-turn is defined between three consecutive residues and contains hydrogen bond between carbonyl oxygen CO of residue *i* and backbone amide NH of residue *i+2* [16]. There are two types of γ-turns, inverse and classic, depending on the torsion angle values of the residue at position *i+1* [17]. The $\phi\psi$ torsion values defining

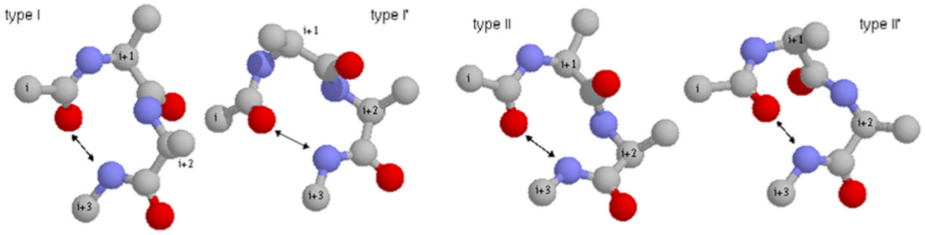


Figure 1.3: A different types of β-turns (I, I', II, II') depending on the value of ϕ and ψ torsion angles of the residues *i+1* and *i+2* [15].

the γ-turn are given in the Table 1.2. It shows that γ-turns are related like mirror images, just like β-turn types I and I', or types II and II'.

Type	ϕ **(i+1)**	ψ **(i+1)**
$\gamma_{classical}$	+75	-65
$\gamma_{inverse}$	-75	+65

Table 1.2: Type of γ-turns and their ideal ϕ and ψ angles. The values of the ϕ and ψ angles are allowed to vary +/- 30° from the ideal turn values.

Classic γ-turns are less common in proteins, but they are responsible for 180° flip of the polypeptide chain. On the other hand, inverse γ-turns are more common, and they tend to introduce a kink in the polypeptide chain [18].

1.2.3 Hydrogen bonds

Besides the peptide bond rigidifying the peptide structure (see Section 1.1), the secondary and tertiary structural elements are also stabilised by forming hydrogen bonds between the residues.

A hydrogen bond is formed when a hydrogen atom is shared between two strong electronegative atoms, usually referred to as atom donor (D) and atom acceptor (A), and binds them together. The hydrogen bond is defined by the distance d between acceptor and donor atoms, $d = D - H...A$ which is typically in range 1.6 - 2.5 Å and the angle enclosed by acceptor and donor atoms $\theta = D - H...A$, which value is between 90° and 180° [19]. The example of the hydrogen bond formed between two alanine dipeptides is shown in Figure 1.4.

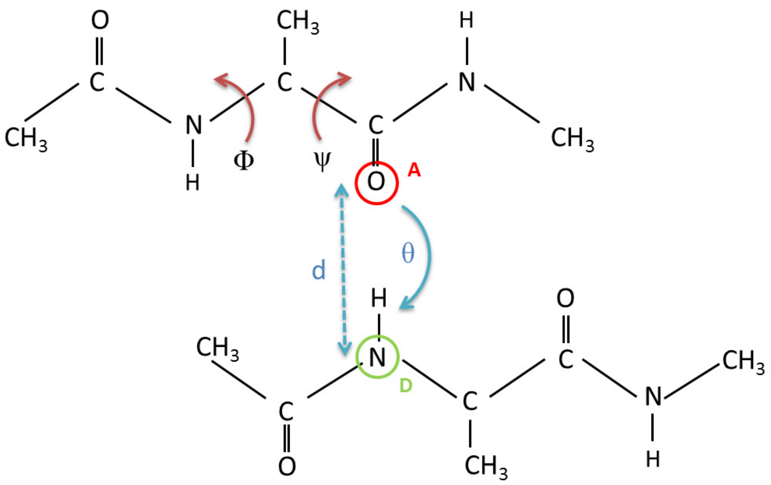


Figure 1.4: The hydrogen bond formed between two alanine dipeptides. d is the distance between donor D and acceptor A atoms, and θ is the angle between them. The ϕ and ψ are torsion angles used to describe peptide backbone conformation.

1.3 Cyclic peptides

Several factors affecting peptide conformational flexibility, such as the peptide bond, secondary structure elements and hydrogen bonds have already been described. Another feature that restricts peptide conformational flexibility is cyclisation [20].

Cyclic peptide is defined as a polypeptide chain in which two parts are covalently linked to make a cyclic motif. The classification covers both natural and synthetically synthesized cyclic peptides. Based on the bond type between the amino acids in the cyclic part of the structure, cyclic peptides can be classified as either *homodetic* (only peptide bond is present) or *heterodetic* (other functional groups but peptide bond are used to connect amino acids) [21]. There are different ways in which cyclic peptide can be formed:

- *head-to-tail* (homodetic): cyclic part is formed between the N-terminal amino group and C-terminal carboxyl group making peptide bond [22]

- *side chain to side chain* (homodetic or heterodetic): the bond is formed by the side chains of different amino acids [23]

- *head to side chain* or *side chain to tail* (heterodetic): a cyclic part is formed by the N- or C- terminal connected to the side chain functional group [24]

Naturally occurring cyclic peptides are most often formed by *head-to-*

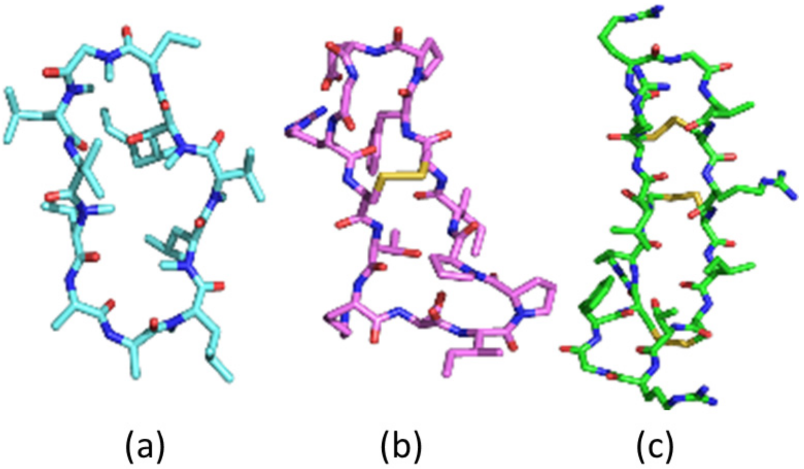


Figure 1.5: The examples of cyclic peptides: (a) Cyclosporin A (PDB ID: 1IKF), (b) Sunflower trypsin inhibitor (STF-1) (PDB ID: 1SFI), (c) Theta defensin 1 (RTD-1) (PDB ID: 2LYF)

tail cyclisation [25], with the well known example cyclosporin A [26] (Figure 1.5 (a)). This category also includes peptides connected by disulfide bond between cysteine residues [27]. The examples are sunflower trypsin inhibitor (STF-1) from sunflower seeds connected by one disulfide bond [28] (Figure 1.5 (b)) or theta defensin connected by three disulfide bonds expressed only in macaques and Old World monkeys [29] (Figure 1.5 (c)). This category also involves plant peptides *cyclotides* in which cyclic part is formed via *head-to-tail* cyclisation, additionally strengthened by three disulfide bonds [30].

Although there exist biologically active cyclic dipeptides - known as diketopiperazines [31], particularly interesting are small cyclic peptides with three to six ring residues [32]. Despite being small in size, they can show a variety of conformational flexibility, including the occurrence of *cis* peptide bond [33]. While in linear peptides, *cis* isomerisation is present mostly in peptide bond involving Proline residue [34], in cyclic peptides, the *cis/trans* ratio was shown to be dependent on the ring size with smaller ring size correlated with the higher occupation of the *cis* peptide bond [35]. This is probably due to the high *cis/trans* energy barrier (approx 20 kcal/mol), while it is lowered (approx 15 kcal/mol) in Xaa-Pro amide bond (Xaa - any amino acid) [36].

Other factors limiting the conformational diversity of cyclic peptide involve the formation of intramolecular hydrogen bonds, or they become conformational restrained by forming beta turns [4]. Moreover, it was found that cyclisation promotes secondary structure β-turn formation in peptides [37], while internal hydrogen bonding promotes passive membrane permeation [38].

Thus, the prediction of the cyclic peptide structure remains challenging task because of the different factors affecting the conformational flexibility [41], and computational methods have emerged as a promising tool to elucidate their detailed conformational diversity [39, 40]. For example, using enhanced sampling computational method short cyclic tetrapeptides were found to adopt several interchanging conformations [36].

1.3.1 Peptide hormones

A special class of cyclic peptides are those which act as peptide hormones [33]. There are a few cyclic peptide hormones expressed in humans - all creating a cyclic part by disulfide bond between two cysteine residues, which differ by the length of the ring part:

- 6-membered ring: Arginine-Vasopressin, Oxytocin, Urotensin II, Urotensin Related Peptide
- 10-membered ring: Melanin-concentrating hormone
- 12-membered ring: Somatostatin-14

Cyclic peptide hormones are secreted by multiple endocrine organs, such as hypothalamus, pituitary gland, as well as different tissues such as heart, pancreas, kidney etc. After being released into blood stream, hormones bind to the specific plasma membrane receptors, initiating signal transduction via secondary messenger system connected to receptors, and subsequently generate specific metabolic responses [42].

1.3.2 Receptor binding mechanism

The traditional *lock-and-key hypothesis* assumes that biomolecules adopt a single conformation which is also their bioactive conformation [43]. This mechanism, introduced by Fischer in 1894, was then extended to the *induced fit* mechanism which suggests that an enzyme changes its shape upon ligand binding to receptor [44]. This model is also known as gloves fitting the hand model. Therefore, unlike the lock-and-key mechanism which describes receptor and ligand as rigid molecules that need to fit perfectly to trigger reactions, induced fit mechanism describes that the enzyme active site adjusts its shape to fit substrate.

However, especially relevant to this work are cyclic peptides binding to G-protein coupled receptors (GPCRs), and the associated binding mechanism. GPCR receptors are the largest family of the cell surface receptors in our body. They are made of 7 transmembrane (TM) helices embedded in a membrane connected with intracellular and extracellular loops. When a ligand binds, it causes conformational change in GPCR resulting in cascade reactions leading to different biological responses [45].

It was long thought that GPCRs act like switches; agonist binding activates GPCR signalling while antagonist binding prevents GPCR activation. The proposed mechanism for agonist/antagonist activity includes interaction with the cell surface [46] or the extracellular loops first [47], and then insertion of the peptide ligand into the receptor binding pocket.

However, experimental evidence emerged revealing that distinct GPCR conformational subsets are activated when different ligands bind, and based on that, each triggers different cascade pathways [48]. This mode of action has been referred to as *biased agonism*, and has also been observed for other receptors too, such as the CCR7 receptor [49].

For example, biased agonism is suggested as a mode of action for two cyclic hormone peptides, Urotensin II and Urotensin Related Peptide which bind to the same receptor, but exert different actions [50]. However, the detailed description of the receptor activation via the biased agonism mechanism remains unclear [51].

1.3.3 Biological activity

Finally, the peptide structure is only relevant in terms of their biological activity. A peptide is considered *bioactive* if it shows an effect on bodily functions. Peptides are involved in many biological processes - they can act like hormones or drugs, and their activity ranges from antimicrobial, anticancer to diuretic, anti-inflammatory, cytotoxic etc. [52]. The cyclic peptides also show a range of biological activity, with Table 1.3 giving some of the examples of the cyclic peptides with associated biological activity.

Of particular interest in this study are cyclic hormone peptides, Vasopressin and Oxytocin (Table 1.3), which function by binding to the G-protein coupled

Cyclic Peptide	Function	Reference
Cyclosporine A	antifungal, anti-inflammatory	[53]
Cyclomarine A	anti-tubercolosis, anti-malaria	[54]
Oxytocin	uterus contractions during labour	[55]
Theta Defensin	antimicrobial	[56]
Valinomicyn	cytotoxic	[57]
Vasopressin	signal transduction, antidiuretic activity	[58]

Table 1.3: A few examples of the functional diversity of bioactive cyclic peptides.

receptors.

1.3.4 Cyclic peptides in drug discovery

Cyclic peptides are particularly interesting in terms of drug design because it was found that they are modulators of protein-protein interactions [59]. The main advantage for the use of these peptides as drug molecules is that they are less toxic and have good binding affinity compared to small molecules [60]. However, there are some disadvantages compared to small molecules. Peptides are poorly orally absorbed and are prone to enzymatic degradation in the gastrointestinal tract [61]. However, their degradation products are less toxic to our organism compared to small molecules because the peptides are made of amino acids which are essential for body function. Moreover, N-methylation was found to improve intestinal permeability for some cyclic hexapeptides [62]. Furthermore, it is assumed for cyclic peptides, due to their ability to form intramolecular hydrogen bonds (leading to a reduction of the hydrophilic surface), that this structural characteristic may facilitate membrane crossing. The Cyclosporin A is an example of a membrane permeable cyclic peptide that crosses the membrane by hiding its hydrogen bonds by forming intramolecular hydrogen bonds [38].

Another advantage of cyclic peptides as drug molecules includes high specificity to their targets and high potency which makes them attractive in terms of peptide-based drug research [63]. Special interest is seen for GPCR-targeted drugs which make 30-50 % of the global market [64]. Many of these drugs belong to the class of peptide hormones, and of these, cyclic peptides are the most interesting because of their low degradation and high bioavailability. There are a few synthetic peptides derived from the cyclic hormone peptides studied in this work. Some of the Vasopressin peptide derivatives include argipressin, desmopressin acetate, lypressin, all acting as therapeutics for diabetes insipidus [65]. Oxytocin derivatives on the drug market are carbetocin and atosiban acetate used to stop after Caeserean bleeding and premature contractions, respectively [65].

1.4 Motivation for the study

As it was shown, the peptides are dynamic molecules with versatile biological activity. Although they are small in size, they are rather flexible which helps them perform different functions. In the focus of this thesis was elucidating the conformational space of the four small cyclic peptides known to be involved in multiple physiological functions in our body. The studied peptides are known to act as ligands to GPCR receptors, which makes them attractive target for the drug design.

Since structure and function are closely related to each other, knowing their structural dynamics may contribute to understand their bioactivity. The characterisation of their conformational space was explored using a combination of methods. Results obtained using enhanced sampling methods, DFT chemical shift calculations and NMR chemical shifts were used to gain knowledge of their conformational equilibrium in solution. The aim of the thesis was therefore to understand the conformational ensemble of the four cyclic peptides in solution in order to explore to what extent their unbound conformational dynamics is predictive of their bound crystallographic conformational state.

The small cyclic hormone peptides studied in this work can also be considered as belonging to the class of intrinsically disordered peptides, whose structure is commonly determined using NMR experiment. The overview of the NMR technique is given in Chapter 2. Next, the literature review, simulation setup and results are given separately for each of the four peptides in the subsequent four chapters. Finally, the comparison between the conformational ensembles of all studied peptides, together with the comparison between the performance of the enhanced sampling methods used to explore their conformational flexibility is given in the last chapter.

Chapter 2

Intrinsically disordered peptides

It was long thought that protein structure is well defined from its amino acid composition. However, more recently proteins with rather flexible yet biologically active structures were discovered, but were given different names, such as natively denatured, intrinsically unstructured, intrinsically unfolded etc. so the classical paradigm was hard to break [66].

Furthermore, it was recognised that this sequence - structure - function paradigm is not true for all the proteins encoded in the genome, and that there are naturally flexible proteins with more than 30 % presence in the eukaryotic proteome [67]. Therefore, the term “intrinsically disordered ”protein (IDP) has emerged [67], referring to a corresponding protein (or protein region) that is biologically active yet exists in an ensemble of flexible conformations.

Sequence signature. Since protein function is related to sequence, the sequence composition of the intrinsically disordered proteins was analysed in the same way. It was revealed that IDPs sequence is enriched in the hydrophobic (Trp, Phe, Tyr, Leu) and charged residues (Lys, Arg, His, Asp, Glu) [68]. This sequence composition was recognised to be correlated with the low protein compactness promoting flexible secondary structures, such as coils or turns, rather than the compact, globular protein conformation [69].

However, there is a considerable interest in revealing how these unstructured proteins perform their biological functions. The first step in discovering this is knowing the mechanism of the binding to receptor.

Receptor binding. An IDP binding to the target receptor has been described as either gaining the structural compactness known as “folding upon binding” or maintaining its solution structure known as “folding before binding” [70].

In the context of IDPs, folding upon binding means that the conformational rearrangement is occurring in such a way that an unfolded IDP becomes structured once bound to the complex [70]. This mechanism is also known as “coupled folding and binding” or “induced fit [70, 71].

However, an IDP can also preserve the structure already pre-existing in solution in a binding mechanism known as “conformational selection” or “folding before binding” [72]. Conformational selection implies that the bound conformation was already formed in solution, and that IDP preserves this conformation upon binding. Experimental evidence for this mechanism was found for the nuclear coactivator binding domain [73] and redox switch CP12 protein [74], and came from combined NMR and X-ray analyses. However, it should be emphasised that ligand recognition is not exclusive to one of these two mechanisms but can also be a combination [75].

Another proposed mechanism of action is the so called fly-casting mechanism where the unfolded state binds weakly at a relatively large distance followed by folding as the peptide approaches the binding site [76]. Another interesting interaction mechanism includes IDPs binding to the same receptor in so called many-to-one mode, or an IDP binding to different receptors (one-to-many mechanism) [70].

To summarise, IDPs are a new subclass of biomolecules which have only recently been recognised. The classification includes both peptides or proteins as a whole, or only specific protein regions. IDPs are widely expressed in the human genome, and have a variety of functions, from transcription and post translational regulation to signalling [77], and involvement in neurodegenerative diseases [78]. However, in order to perform their function, they need to bind to a particular receptor to start the signalling pathway. A few binding mechanisms were reviewed in this section, indicating that IDPs have no single preferred binding mode, which could be due to their inherent structural diversity. Therefore, the knowledge of their structural ensemble may help in understanding the molecular mechanisms responsible for their function.

2.1 Conformational ensemble of IDPs

The knowledge of the molecule 3D structure gives an insight into its biological functions. In a cell, molecules are constantly on the move, and consequently may change their structure as a result of their dynamics to match their function. Particularly challenging is the characterisation of IDP structural ensemble because they are believed to adopt a range of structural substates.

The most common experimental technique for determining IDP structure is NMR spectroscopy [79]. Compared to X-ray diffraction where only a single structure is obtained, NMR captures peptide dynamics. Moreover, the advantage of NMR over other experimental methods is that it gives the information about the peptide dynamics in a solution state, which is close to the natural physiological environment. NMR can also provide the peptide ensemble conformation in contact with SDS micelles which mimic the cell membrane environment.

However, structural complexity can sometimes prove hard to experimentally characterise with NMR due to time and ensemble averaged signals being translated into structural features. Ensemble average means that the obtained ensemble is averaged over many microstates, and time ensemble means that measured parameters are averaged over a certain period of time [80]. For example, in NMR the motion of typically $10^{14} - 10^{17}$ molecules in the test tube is fitted to averaged experimental data.

Despite that, NMR can still provide valuable information about peptide structural diversity. The information gathered from NMR is contained in the different measured observables: *chemical shifts* provide with secondary structure content of the molecule,; *spin-spin or J couplings* report on backbone dihedral angles; the *nuclear Overhauser effect (NOE)* specifically provides distance information, and *residual dipolar couplings (RDC)* report on the orientation of the spatially distant parts of the protein [81].

In the following sections, more detailed descriptions of each of the NMR observables is given and how each is transferred into the peptide 3D structure information.

2.1.1 Nuclear Magnetic Resonance

Nuclear Magnetic Resonance (NMR) is an experimental technique commonly used to determine structure of biomolecules. The experimental instrument consists of a magnet which produces uniform, intense magnetic field, a signal amplifier, detector and receiver. A typical NMR superconducting magnet is producing a magnetic field of 10 T or more [82]. The reason of using such strong magnetic fields is two fold; stronger magnetic field intensifies the transitions between the energy levels, and simplifies the appearance of certain spectra lines. The sample is kept in between the magnet and is exposed to the flashes of the radio frequency (RF) short impulses which causes certain nuclei to excite and the signal is measured.

The measured phenomena arises from the fact that electrons and nuclei possess the intrinsic property called spin. When a magnetic field is applied to the electron, it can adopt two spin orientations, $m_s = \pm 1/2$. On the other hand, nuclear spin can adopt $2I + 1$ orientations, where I is nuclear spin quantum number, which depends on the particular nuclear under consideration.

Each electron spin state has its associated energy given by

$$E_{m_S} = -g_e \gamma \hbar B m_S \tag{2.1}$$

where g_e is a g-value for the electron which has a value of 2.0023 for the free electron, B is magnetic field, $\hbar = h/2\pi$ is reduced Planck constant, and γ is the magnetogyric ratio.

$$\gamma = -\frac{e}{2m_e} \tag{2.2}$$

where m_e is the electron mass, and e is the magnitude of electron charge. It is common to express energy in terms of *Bohr magneton* which is defined as

$$\mu_B = \frac{e\hbar}{2m_e} \tag{2.3}$$

Substituting this into the energy term (Equation 2.1), a new expression for the energy of the spin state is obtained

$$E_{m_S} = g_e \mu_B B m_S. \tag{2.4}$$

The electron spin states are usually denoted as $m_S = 1/2 = \alpha$ and $m_S = -1/2 = \beta$. The α state is at higher energy than the β state.

The energy difference between two spin states in then given as

$$\Delta E = E_\alpha - E_\beta = \frac{1}{2} g_e \mu_B B - (-\frac{1}{2}) g_e \mu_B B = g_e \mu_B B \tag{2.5}$$

The same set of equations may be applied to the nuclear spin system. However, in the magnetic field B, the nuclear spin can have $2I + 1$ orientations. The energy of each level is given as

$$E_{m_I} = \gamma_N \hbar B m_I \tag{2.6}$$

where γ_N is the nuclear magnetogyric ration, m_I is the nuclear spin angular moment taking values $m_I = I, I-1, I-2.... -I$, $\hbar$ reduced Planck constant and B is magnetic field. The energy term is usually written in terms of nuclear megneton μ_N

$$\mu_N = \frac{e\hbar}{2m_P} \tag{2.7}$$

where m_P is the proton mass and e is the magnitude of proton charge. In that case, the energy of the nucleus is expressed as

$$E_{m_I} = -g_I \mu_N B m_I \tag{2.8}$$

For the nuclear spin system I=1/2, the energy level $m_I = -1/2 = \beta$ is at higher energy level than $m_I = 1/2 = \alpha$.

If we look at the electron and nuclear spin systems with classical analogy, then they can be considered as tiny magnets. In the presence of magnetic field B, they are not perfectly aligned along the B axis, but they orbit around it with angular frequency known as *Larmor frequency* given by

$$\omega = \gamma B \tag{2.9}$$

which only depends on the magnetogyric ratio γ, and magnetic field B.

When a radio frequency (RF) pulse is applied with appropriate energy (equal to the difference in energies of the two levels), transitions between the two energy levels will be induced. This suggests that the *resonance absorption* happens when the precession frequency matches that of the applied radio frequency field, and as a result the intensity on the spectrum is recorded.

2.1.1.1 Chemical shifts

The chemical shifts phenomenon comes from the nuclear spin coming into contact with the surrounding magnetic field. The magnetic field felt on the nucleus comes from both the externally applied magnetic field and from the magnetic field produced by the surrounding electrons. The magnetic field experienced from the local environment is expressed as

$$B_{add} = \sigma B \tag{2.10}$$

where σ is the shielding constant which can have positive or negative value depending on whether the induced field adds or subtracts from the applied field. This local electron magnetic field ”shields” the nucleus from the full force of external magnetic field. In that case, the total magnetic field experienced by the nucleus becomes

$$B_{loc} = B + B_{add} = (1 - \sigma)B \tag{2.11}$$

and the resonance condition equals

$$\nu = \frac{\gamma_N B_{loc}}{2\pi} = \frac{\gamma_N}{2\pi}(1 - \sigma)B \tag{2.12}$$

Having introduced the shielding constant, the chemical shift of the nucleus is defined as the difference between its resonance frequency and that of a reference standard. The common reference standard in proton resonance is TMS or chemically known as tetramethylsilane $Si(CH_3)_4$. Shielding constants are converted into chemical shifts using this formula

$$\delta = \frac{\nu - \nu^0}{\nu^0} \times 10^6 \tag{2.13}$$

where ν^0 is a resonance frequency of the standard. The chemical shifts are measured in parts per million (ppm).

Chemical shifts are often used to assign secondary structure based on the chemical shift index method [83]. Based on the set of rules, the measured chemical shifts of the particular nuclei are then compared with the reference values, and the protein conformational state can bes assigned as a helix, beta strand or random coil. In addition, the tables of typical ranges for the ^{1}H chemical shifts for certain chemical groups exist that facilitate the assignment of the NMR spectra [82].

2.1.1.2 Spin - spin coupling

There is usually more than a single signal peak appearing for a given chemical shift in an NMR spectrum. The signal is often split around a central chemical shift. This phenomenon comes as a result of the neighbouring spin spin interactions in the system. The strength of interaction is expressed in terms of the *spin-spin coupling constant J* and is measured in Hz. The spin spin interaction is an intrinsic property of the molecule and does not depend on the magnitude of applied magnetic field.

In the case of a ***one proton system***, a nucleus of $I = 1/2$ spin will have $2I + 1 = 2$ orientations. These orientations are of equal energy so in the absence of external magnetic field, there is no energy state splitting.

However, when a magnetic field is applied, the energy levels split as is observed in Figure 2.1. The lower energy state is more populated than the higher energy state. This is given by the Boltzmann distribution. If the system is irradiated at the frequency that matches the energy difference between the energy states, then we get *resonant conditions* and the population of the energy states equalizes. This is observed as the appearance of the signal on the spectrum (Figure 2.1).

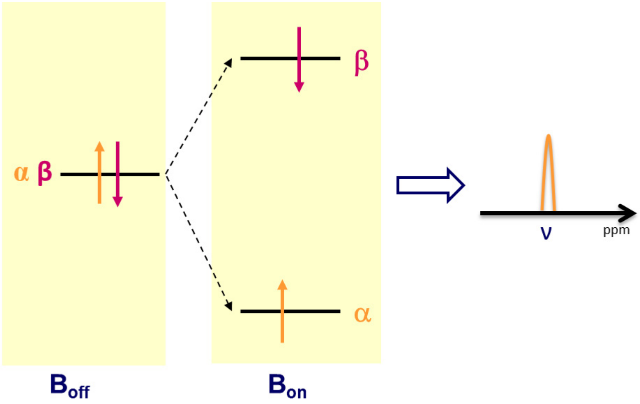


Figure 2.1: In the absence of the external magnetic field B_{off}, the spins are occupying the same energy state. However, when the external magnetic field B_{on} is applied, the energy levels split what is observed as the appearance of the signal line on the spectrum.

In case of a ***two nuclei AX system*** (letters far apart mean that the associated nuclei chemical shifts are very different), there are four energy levels possible because each nucleus has two possible spin energy states (Figure 2.2).

If we first consider proton A, and it changes its spin state from α to β, X nucleus remains in its spin state, which can be α or β. This is observed as splitting of the signal separated by J because of the two transitions possible. The same applies to the X nucleus, which can change its spin state from α to β, but the A nucleus remains in one of its spin states, α or β. There are two transitions possible, and this is also observed as the signal splitting differing in frequency by J. The mechanism is shown on the Figure 2.2.

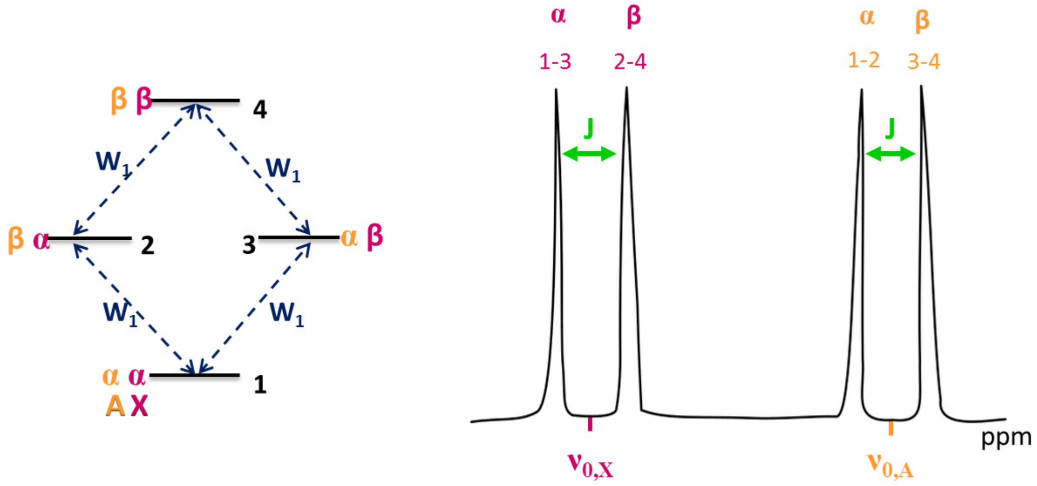
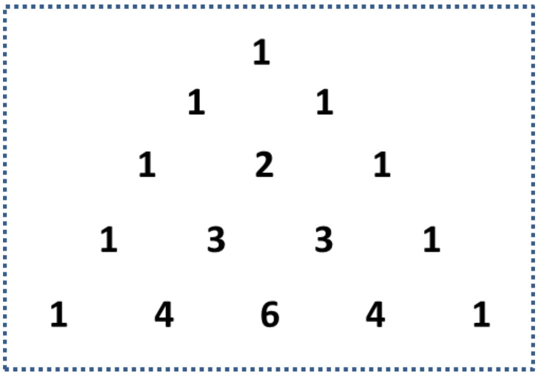


Figure 2.2: Two proton system occupies four energy levels. When A nucleus changes its spin state from α to β, the X nucleus remains in on of its spin states, α or β. This corresponds to 1-2 and 3-4 transitions. The same applies to X nucleus; it can change its spin state from α to β, with A nucleus remaining in its spin state, which is either α or β. This transition is observed as 1-3 and 2-4 transition on the plot. The J coupling is observed as signal splitting on the spectrum, shown on the right.

The splitting of the signal follows a simple rule where n neighbouring magnetically equivalent nuclei split the signal into n+1 multiplets with intensities following Pascal's triangle rule (Figure 2.3). Protons that are separated by four or more bonds do not couple.



1
1 1
1 2 1
1 3 3 1
1 4 6 4 1

Figure 2.3: Pascal's triangle is used to predict the intensity of the lines in the NMR spectrum.

J-coupling constant is related to the electronic structure, geometry, and conformation of a molecule. However, specially important is the dependence of the coupling constant on the torsion angle defined by three bonds. This dependence is explained by the **Karplus equation** [84]

$$^{3}J(H,H) = A + Bcos\phi + Ccos2\phi \tag{2.14}$$

It states that if we know the coupling constant between hydrogen atoms separated by 3 bonds, then the angle between them can be determined using the above equation. It predicts the angle between H-x-x-H atoms, where x is any atom. The above constants A, B, C take the values +7 Hz, -1 Hz and +5 Hz, respectively in case of the H-C-C-H dihedral angle [84]. Their values are empirically or experimentally derived and depend on the substituents involved [85].

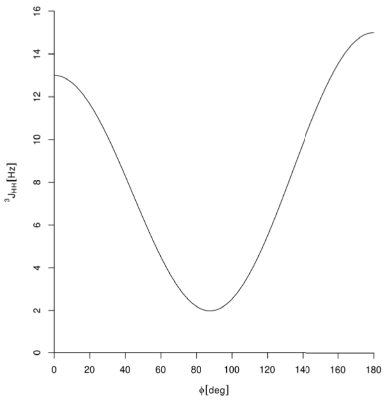


Figure 2.4: Karplus curve shows the dependence of $^{3}J(H,H)$ on the dihedral angle in H-x-x-H system where x is any atom [82].

2.1.1.3 Nuclear Overhauser Enhancement (NOE)

The Nuclear Overhauser Enhancement (NOE) effect is defined as the transfer of the nuclear spin polarisation from one nucleus to another via the process called *cross relaxation* [86]. In the equilibrium state, the lower energy state $m_I = 1/2 = \alpha$ is more populated than the higher energy state $m_I = -1/2 = \beta$. When a radio frequency (RF) pulse is applied to the system, both energy levels become equally populated what is called *saturation*, and after that the *relaxation* process occurs.

We can now consider two spin system AX interacting through dipole-dipole interaction. A coupled two nuclei system adopts four energy states (Figure 2.5). In the lowest energy state (denoted as 1), the spins are in α orientation while in the highest energy state both spins are in β orientation (energy level 4).

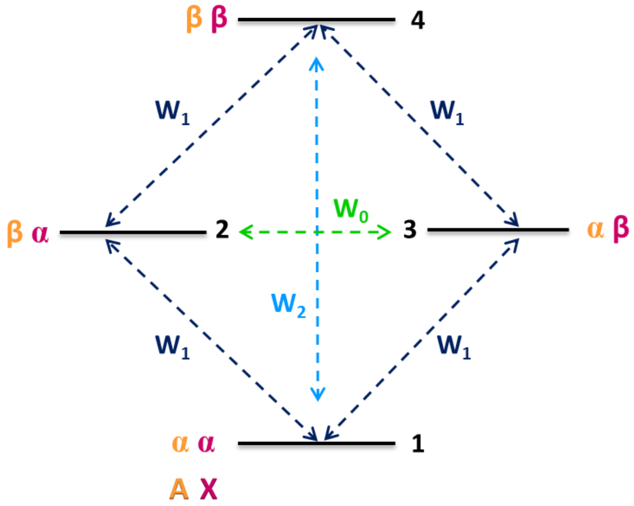


Figure 2.5: Energy levels of the two spin system AX interacting through dipole-dipole interaction.

The normal signal intensities of the A and X resonances are determined by the population difference between lower and higher energy states in a spin transition. In the NOE experiment, the X nucleus is irradiated with double resonant frequency, and the change of the signal intensity is monitored for the other nucleus A.

When the X resonance is irradiated, then the population difference between the X energy spin states decreases in a saturation process. On the example in Figure 2.5, it means that the population of the energy levels 3 and 4 has increased, while the population of the energy levels 1 and 2 has decreased.

A saturation of the energy levels is then followed by the relaxation process. There are two relaxation processes possible; W_2 also known as double quantum transition, and W_0 known as zero quantum transition.

If the relaxation happens from $\beta\beta$ to $\alpha\alpha$ spin energy states, labelled as W_2 in Figure 2.5, then an enhancement of the A proton signal intensity is observed, and this is called positive NOE. Another relaxation process W_0 also moves the irradiated system back to equilibrium, but in this relaxation mechanism the A signal intensity is decreased, what is referred to as negative NOE.

A difference between the W_2 and W_0 relaxation rates in comparison with all possible relaxation rates is related to the change in signal intensity, and it is reported in terms of parameter η. The effect is observed by comparing the signal intensities I with the normal intensity I_0 measured in the absence of double radiation.

$$\eta = \frac{I - I_0}{I_0} \tag{2.15}$$

It was found that the value of the η depends on the interproton distances as r^{-6}. This is the most important usage of NOE because from the signal spectra, it can be built up a picture of the molecular system conformation by identifying nuclei that are spatially close together [82].

A typical NMR experiment consists first of identifying all the nuclear resonances in the molecule, and then saturating them one by one measuring the enhancement in the signal. In this way, it is possible to relate for example if two parts of polypeptide chain came close together upon folding.

2.1.1.4 Residual Dipolar Coupling

When two nuclei are coupled together, there may be an additional effect influenced by the applied magnetic field which gives rise to the dipolar interactions manifested as *Residual Dipolar Coupling (RDC)* [87].

All other NMR methods give local information about the molecular system; the distance between spatially close nuclei or the dihedral angle between them. However, the long distance information was not available before the discovery of RDC.

If the two nuclei that are coupled together are put in an external magnetic field B_0, then the vector between them can be defined relative to the orientation of the external magnetic field. The RDC is related to interproton distance r and the angle between the bond and magnetic field θ as

$$D \approx (3cos^2\theta - 1)r^{-3} \tag{2.16}$$

If the vector connecting nuclei A and X is parallel to the field B_0, the coupling is at its strongest.

The RDC is measured in an anisotropic media because in an isotropic medium molecules tumble quickly, and $3cos^2\theta - 1$ averages to zero. To exhibit RDC, slow tumbling needs to be obtained, or preferential ordering along a particular direction. Different aligning media have been used for aligning purposes, such as phospholipid bicelles, magnetically oriented viruses or polyacrylamide gels, to name a few [88].

In the NMR experiment, RDC measurement consists of doing two experiments in parallel, one with and one without the presence of an aligning medium. After an

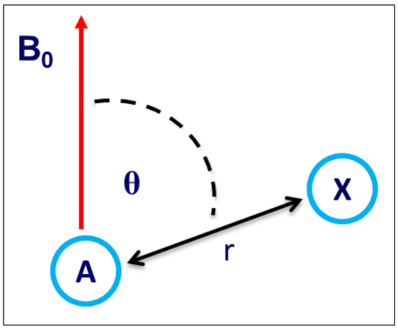


Figure 2.6: Residual Dipolar Coupling measures the relative orientation θ between two nuclei, A and X, and the magnetic field B_0.

identical NMR pulse sequence is applied to both samples,*J-coupling constant* and $J + D$ values are measured. The difference yields the dipolar coupling D which reports on the orientation between the nuclei in the solvent [89].

2.1.1.5 Summary

NMR experiments provide with a large amount of experimental data which may be used to obtain the protein conformational equilibrium *in solution*. The experiment is usually performed at pH 6 - 7 to account for cellular physiological conditions. Usually when doing the experiment, chemical shifts are first measured which report on the secondary structure of the system. Then, spin - spin interactions which strongly depend on the local magnetic field experienced by each of nuclei, can be measured. We distinguish between spin *through bond* coupling which is reflected in J-coupling values, or as a *through space* interaction manifested as NOEs.

NOEs measure the interaction of the spins that are at the distance less than 5 - 6 Åto each other *in space*. The NOE intensities are related to interproton distances as r^{-6}. As a result, NOE spectrum relates the chemical groups that are close in space.

Next, J-coupling provides the value of the torsion angle between nuclear spins connected *via three bonds*. These values have been widely used for the conformational analysis of the biomolecules. While J-coupling reflects the local conformation of the molecule, RDC is used to relate spatially distant parts of the protein to each other.

The timescale of the dynamical motion accessible to NMR techniques covers a range between very fast ($< 10\mu s$) and very slow conformational dynamics (> 10 ns) [90]. Fast conformational isomerisation is usually observed as a single signal with averaged chemical shift values reporting the ensemble average conformation, while motions on the slow NMR timescale resolve separate signals for each conformation [91]. The IDPs in the focus of this thesis were often reported as a single structure or unstructured in solution phase, suggesting that their conformational interconversion is happening on the timescale that is too fast for NMR to detect as separate signals.

Chapter 3

Methods

Computational chemistry has increasingly been used to study biological systems. This has primarily become possible due to increases in the available computational power and advances in the simulation methodology. Biomolecules are dynamic systems which can, in order to perform different functions, undergo the conformational changes due to interaction with receptor or acting as signalling molecules etc [92].

Although X-ray crystallography provides atomic resolution for the system, it is now widely accepted that proteins undergo structural changes to maintain their function, which implies that the crystallographic picture of the system is not necessary the only one in terms of structural arrangement [93]. On the other hand, other experimental techniques, such as spectroscopic methods, lack the atomic level accuracy.

An alternative approach to get detailed structural and dynamic information about biological systems is by using computational methods, which study system either at the atomic (Molecular Dynamics) or at subatomic level (Quantum Chemistry).

The most accurate picture of system behaviour is provided by quantum mechanics (QM); however a key challenge is the amount of computer power used due to the scaling as N^4 or N^3 with system size [94], which remains infeasible for biomolecular systems. Because of that, Molecular Mechanics is used instead, which makes use of forcefields to model interactions within the system. However, to access long timescales in Molecular Dynamics we need advanced sampling methodologies, for example Metadynamics or Replica Exchange Molecular Dynamics, which are able to overcome the problem of insufficient sampling in different ways. Therefore, this chapter provides an overview of the theoretical principles behind the methods used in this work, which range from QM to enhanced sampling molecular dynamics.

3.1 Introduction to Quantum Mechanics

In classical mechanics, based on the current knowledge of the system, we can predict its behaviour in the future. However, some experimental observations before the 1900s could not be explained with classical physics. The experiments, such as photoelectric effect and black-body radiation, could only be explained by introducing wave-particle duality and energy quantisation [95]. These concepts set the fundamentals of QM.

First, the concept of a **wave function** needs to be introduced. At any instance in time t, the physics of an electron (or nucleus) can be described with the wave function $\psi(r,t)$ [96]. The wave function is a completely quantum concept and there is no classical analogy. The most intuitive interpretation of it was given by Max Born who gave the probabilistic interpretation that states that the square modulus of the wave function is a probability density

$$P(r) = \psi^{*}\psi = |\psi|^{2} \tag{3.1}$$

Another quantum concept that needs to be introduced is spin. **Spin** is an intrinsic property of the particles. The value of spin is defined by the quantum number. Electrons have one half integer spin while different nuclei have different spin values. In a classical analogy, particles with spin act like tiny magnets in magnetic field.

Finally, we will introduce the time independent **Schrödinger equation** which gives a way of calculating the energy of the system

$$\hat{H}\psi = E\psi \tag{3.2}$$

The equation states that Hamiltonian operator $\hat{H}$ acts on wave function ψ and returns the energy of the system E multiplied by the wave function. The operator is a mathematical entity that acts like a function. In QM, the Hamiltonian operator is defined as a sum of operators associated with kinetic and potential energy

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r}) \tag{3.3}$$

where $\hbar$ is Planck's constant over 2π, m is the mass of the particle, ∇ is the Laplacian operator $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. The Schrödinger equation can be exactly solved only for the one-electron system of a hydrogen atom, which is composed of one proton and one electron. Using the *Born Oppenheimer approximation* the motion of the nucleus and electron can be separated. Since the mass of the atom is localised in the nucleus, we can get its position using classical Newtonian equation. The electron position is obtained using the Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} + V(x)\psi = E\psi \tag{3.4}$$

which describes a particle of mass m with energy E.

However, there is no exact solution to Schrödinger equation for two electron system, the wave function contains the electron - electron repulsion term which is only possible to solve approximately. Different ways of solving this problem have been introduced, and here two of them will be introduced, Hartree-Fock and Density Functional Theory.

3.1.1 Hartree-Fock Theory

The Hartree-Fock (HF) theory [97, 98] is based on solving the many electron problem as a series of integrals over one electron at a time. The method is also referred to as the Hartree-Fock Self Consistent Field (SCF). It is based around the idea that each electron is moving in the electromagnetic field of the nuclei and it only encounters the other electrons of the system as an averaged effect.

In this method, the position of the electron is approximated by different wave functions, which are then used to calculate the average potential felt by each electron. These potentials are then used to calculate new orbitals (wave functions). The process is iterated until the individual wave functions reproduce the average potential used to calculate them.

Since HF theory takes the electron-electron interactions into account in an averaged way, ignoring correlation, it results in HF overestimating the true ground state energy of the system E_0. The difference between E_0 and E_{HF}, the HF energy of the system, is defined as the *correlation energy* of the system [99]. It accounts for the dynamic correlation (electron repulsion) and static correlation (a single wave function is not always adequate to describe a molecular state).

3.1.2 Density Functional Theory

Instead of dealing with many-electron wave functions, an alternative approach is to deal with *electron charge density*. The charge density $n(\mathbf{r})$ is a much simpler description of the electronic component of the system, and Hohenberg and Kohn showed that ground state energy is defined by the electron charge density [100].

In principle, the total energy of the system can be written as the sum of energy functionals

$$E[n(\mathbf{r})] = T[n(\mathbf{r})] + V[n(\mathbf{r})] + U[n(\mathbf{r})] \tag{3.5}$$

where T - kinetic energy of the electrons, V - electron nucleus interaction, and U - electron-electron term. By definition, a functional is a function that acts on a function and returns a scalar value, in this case energy. The only known term in the Equation 3.5 is electron nucleus interaction V, while the electron correlation functional U and kinetic energy functional T are unknown. The most widely used way to calculate T and U is by using **Kohn-Sham formalism** [101] defined as

$$T + U = T_0 + U_0 + U_{xc} \tag{3.6}$$

where T_0 is a functional that gives the total kinetic energy of a set of N interacting electrons, U_0 is electron-electron repulsion term which treats electrons independently similarly to HF, and U_{xc} is exchange correlation functional. The idea of the Kohn-Sham approach is to work with a system of non-interacting electrons whose density is the same as the system of the interacting electrons. In the Kohn-Sham approach the one electron Schrödinger equation is calculated for an electron moving in an average potential derived from a fictitious system of surrounding interacting electrons. Following this approach, Hamiltonian is actually a sum of one-electron Hamiltonians for non-interacting electrons. The Kohn-Sham Equation is solved in a self-consistent way by first guessing the electron charge density $n(\mathbf{r})$, and then the wave functions (molecular orbitals) of the non-interacting electrons are determined, leading to a better value of the density, which is then used in the next step to calculate orbitals. The process is repeated until convergence is achieved. However, the only unknown term in Kohn-Sham Equation 3.6

remains U_{xc}, the exchange correlation functional. Different approximate methods exist to calculate this term.

In this work, a **hybrid functional** for exchange correlation energy term was used [102]. In particular, the B3LYP version [103]. It includes exact exchange energy from Hartree–Fock theory with exchange and correlation from other sources (ab initio or empirical). The exact exchange energy functional is expressed in terms of the Kohn–Sham orbitals rather than the density, so is termed an implicit density functional. A hybrid exchange-correlation functional is usually constructed as a linear combination of the Hartree–Fock exact exchange functional and estimates of correlation energy functionals.

3.1.3 Basis set

A basis set of functions is required to describe the orbitals for many quantum mechanical methods, so here will be given a brief introduction. A basis set is defined as a set of functions used to represent the electronic wave function. Any function can be written as a linear combination of special basis functions, b_i

$$f(x) = \sum_{i}^{\infty} c_i b_i(x) \tag{3.7}$$

A *complete* basis set is a set that gives all possible forms to f(x). However, for practical reasons in QM computations, only a limited collection of basis functions is used. This being the case, the number of basis sets need to be limited but at the same time provide a good description of the system. The basis set is usually taken as a linear combination of atomic orbitals or plane waves. There are a few types of atomic orbitals developed, such as Slater Type Orbitals (STOs) [104], Gaussian Type Orbitals (GTOs) [105], or numerical atomic orbitals. The most commonly used are GTOs [106].

Different GTO basis sets have been proposed over the years, differing in size. For this reason the concept of the *minimal basis set* was introduced, defined as the smallest required number of atomic orbitals to describe the system of interest. However, a minimal basis set is usually not sufficient to achieve high-level accuracy. This was achieved by, in particular, introducing additional wave functions for the split-valence basis sets to account for valence electrons for a better description of the polarisation effect.

Another improvement of the basis set includes the addition of diffusion functions to basis sets. These functions decay slowly as they move away from the nucleus which captures dispersion and charge transfer.

Typically the notation of the basis sets follows the Pople X-YZ(VW)(+)G(*) form [107]. X is the number of GTOs used within each core electron orbital, the number of digits after the dash indicates whether the basis is double, triple-zeta, etc., with each digit giving the number of GTOs of that basis function. The asterisk denotes the inclusion of polarisation functions, while + represents the inclusion of diffuse functions.

3.2 DFT calculations of NMR chemical shifts

Besides being experimentally measured, chemical shifts can also be calculated for the given biomolecular geometry using empirical or QM based approaches. There are several programs available to predict chemical shifts using empirical approaches [108–110], however their prediction of the chemical shifts even for small organic molecules was proven less accurate [111].

On the other hand, a QM based DFT approach of more accurately calculating chemical shifts was found to assist in the structural assignment of measured NMR chemical shift, facilitating characterisation of reaction intermediates or in studying conformational motions [112]. These examples show that DFT theory can facilitate the experimental chemical shift assignment for a variety of problems.

DFT uses electron density to calculate magnetic properties of nuclei (Section 3.1.2), such as chemical shifts of the given geometry. The electrons surrounding a particular nucleus affect the local nuclear magnetic field, which is known as shielding. Within the DFT method, the nuclear magnetic shielding can be calculated using IGLO (Individual Gauges for Localised Orbitals) [113] or GIAO (Gauge-Invariant Atomic Orbital) [114] techniques, but GIAOs were shown to obtain faster convergence of calculated chemical shieldings [115]. In this work, the standard implementation of GIAO in Gaussian09 [116] was used at the B3LYP/6-31G(d) level of theory to calculate magnetic shielding constants for the given structures.

The calculated isotropic nuclear magnetic shielding constants (σ_H) were converted into chemical shifts (δ) using the regression equation [117]

$$\delta(^{1}H) = -0.9912\sigma_{H} + 32.05 \tag{3.8}$$

The parameters used to convert shielding constants into chemical shift were obtained by calculating the chemical shifts of small organic molecules with DFT calculated magnetic shieldings at the same level of theory as used for the calculation for the peptides studied in this work [117]. The error in the predicted chemical shifts for the training set compared to the experimental data was 0.18 ppm for ^{1}H chemical shifts.

A QM based approach to calculate chemical shifts for a given structure is therefore more accurate than empirical methods because the QM theory provides a more sensitive picture of the local structural arrangement in the molecule. Moreover, this approach includes the solvation effect through the common polarizable continuum model (PCM) representing water as an implicit solvent [118].

3.3 Molecular Dynamics Theory

Molecular Dynamics (MD) studies the time evolution of the system using Newton's laws to predict the position of all the system members at any point in time. Atoms are described as particles with a defined set of parameters representing interactions between them, and their motion is followed over a certain period of time. As a result, an ensemble of configurations is generated providing a dynamical picture of the system. The processes studied by MD range from conformational changes to insertion of peptides into a membrane.

Given the initial coordinates of the studied system, the motion of the system can be followed by solving Newton's equations of motion

$$\mathbf{F} = m\mathbf{a}. \tag{3.9}$$

where **F** is a force, **a** is acceleration and m is mass of the particle. Knowing the position of the particles in the system, it is possible to calculate the force acting on them. The forces are calculated from the gradient of potential energy

$$\mathbf{F} = -\nabla U \tag{3.10}$$

and then positions of atoms are calculated using Newton's second law. Since the forces acting on each atom depend on the position of all other atoms in the system, complex differential equations are generated which cannot be solved analytically. A standard method for solving differential equations is the finite difference technique where positions of atoms are approximated using the Taylor expansion in the small time increment (time step) δt

$$\mathbf{r}(t+\delta t)=\mathbf{r}(t)+\delta t\mathbf{v}(t)+\frac{1}{2}\delta t^2\mathbf{a}(t) \tag{3.11}$$

Several algorithms have been developed to calculate these expansions including the velocity Verlet [119] and leapfrog algorithms [120]. For example, the leapfrog algorithm calculates new position of particles like this

$$\mathbf{r}(t+\delta t)=\mathbf{r}(t)+\frac{1}{2}\delta t\mathbf{v}(t+\frac{1}{2}\delta t) \tag{3.12}$$

Where $\mathbf{r}$ is the position of a particle, $\mathbf{v}$ is the velocity of the particle, which is calculated at $t+\frac{1}{2}\delta t$ as follows

$$\mathbf{v}(t+\delta t)=v(t-\frac{1}{2}\delta t)+\delta t\mathbf{a}(t) \tag{3.13}$$

where $\mathbf{a}$ is the acceleration of the particle. The time evolution of the system is obtained by calculating new positions using the velocity calculated half a time step ahead of the position. The initial velocities are generated randomly to correspond to the desired temperature of the system.

The stability of the integrator depends on the time step length. If the time step is too small, then the method is insufficient due to the high computational cost. If it is too high, the forces will not maintain constant which will result in the loss of energy conservation. Typical time step used is 2 fs which is sufficient to conserve the energy and account for the the fastest motions in the system (hydrogen-containing bond vibrations) which are constrained using SHAKE algorithm [121].

3.3.1 Force field

However, without something to model interactions and calculate the total energy, there is no dynamics. This central part of molecular dynamics is known as the force field. Many families of bimolecular force fields exist today including OPLS [122], GROMOS [123], CHARMM [124] and AMBER [125], the last of which is used in this work. The general form of the total potential energy in the AMBER force field looks like this

$$E_{potential} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{coulomb} + E_{LJ} \tag{3.14}$$

Bonds and angles contributions are represented as balls-on-springs according to Hooke's law

$$E_{bonds} = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 \tag{3.15}$$

$$E_{angles} = \sum_{angles} \frac{k_i}{2}(\Theta_i - \Theta_{i,0})^2 \tag{3.16}$$

where $l_{i,0}$ and $\Theta_{i,0}$ are bond and angle reference values, respectively. The values of these parameters are obtained from vibrational spectroscopy experiments or QM calculations. The third term in equation 3.14 models rotation around bonds

$$E_{torsions} = \sum_{torsions} \frac{V_n}{2}(1 + cos(n\phi - \gamma)). \tag{3.17}$$

This term is parametrised by a Fourier series with Fourier coefficients V_n, dihedral angle ϕ and phase difference γ. The other two terms in the equation represent non-bonded interactions. These are electrostatic and van der Waals interactions. The interaction potential between two charged atoms is given by Coulomb's law

$$E_{coulomb} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}. \tag{3.18}$$

where q_i and q_j are atom partial charges and r_{ij} is the distance between them. ε_0 is electric permittivity. The last term in the potential energy expression is van der Waals term

$$E_{LJ} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right) \tag{3.19}$$

represented by a Lennard-Jones (LJ) potential which accounts for the interactions due to overlap of the electron clouds between two atoms (repulsion) and attraction between induced dipoles which varies as r^{-6}. The LJ potential describes 1-4 interactions (i.e. those between atoms separated by three bonds), but also intermolecular interactions.

In contrast to the Coulomb interactions, LJ interactions decay quickly, therefore the long-range LJ interactions can be treated with cutoff, mostly in the interval of 10-12 Å.

3.3.2 Periodic boundary conditions

Force fields are optimised for treating protein systems in a bulk of solvent. However, molecules at the box surface will experience different kind of forces compared to molecules in the bulk. To overcome this problem, periodic boundary conditions are imposed where a system box is infinitely replicated in all directions. During a simulation, if an atom moves outside the box, its images in the neighbouring boxes are moved in the same direction. Thus, its image atom will move again to the same central box but from the opposite side. However, each atom is allowed to interact only with the closest image of other atoms, and this requirement is called the *minimum image convention*. To avoid an atom interacting with its image, restrictions in term of box size length are imposed which suggest that box size should be at least twice the cutoff value. Typical cutoff value is 12 Å so the box length must be greater than twice the cutoff plus the protein length.

3.3.3 Long range interactions

The interactions in the system are calculated within the cutoff value. However, that causes a discontinuity in the energy term for non-bonded interactions. This is especially problematic for the Coulombic interactions which decay slowly with distance dependence of 1/r, compared to vdW interactions which decay much faster (Equation 3.19). To overcome this issue, the Ewald summation method was introduced which accounts for the electrostatic energy of the system with periodic boundary conditions, where potential coming from the replicated cells is also taken into account [126]. Since it is possible to write any function as a sum of two terms, the same thing was done for the electrostatic energy here. The interaction potential is decomposed into short-range and long-range components.

The idea is that point charges are mapped with Gaussian distributions of the opposite charge. By doing this, the system becomes neutralised and the short range converged contribution to the electrostatic potential is calculated in real space.

The second modification is to superimpose a second set of gaussian charges, this time with the same charge sign as the original point charges and also centred on the point charges. This is done to recover the original system, and this term accounts for the long range interactions. It is calculated in reciprocal space. The last term is the self energy correction. Following this approach, a total electrostatic contribution to the potential energy becomes

$$E_{el} = \sum_{|\mathbf{n}|=0} \frac{q_i q_j}{4\pi\varepsilon_0} \frac{erfc(\alpha|\mathbf{r_{ij}} + \mathbf{n}|)}{|\mathbf{r_{ij}} + \mathbf{n}|} + \frac{4\pi}{L^3} \sum_{k \neq 0} \frac{q_i q_j}{4\pi\varepsilon_0} exp(\frac{-k^2}{4\alpha^2}) cos(\mathbf{k} \cdot \mathbf{r_{ij}}) - \frac{\alpha}{\sqrt{\pi}} \sum_{k=1}^{N} \frac{q_k^2}{4\pi\varepsilon_0} \tag{3.20}$$

where α is Ewald convergence parameter, $erfc(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty exp(-t^2)dt$ is the error function and $k = \frac{2\pi\mathbf{n}}{L^2}$ where $\mathbf{n}$ is cell coordinate vector and L is cell dimension. The α parameter is chosen in a way to optimize the convergence to 0 of the error function erfc(x).

However, the calculation of the long range interactions is computationally expensive using Ewald summation as it scales with the square of the number of particles (N). In this work, the particle mesh Ewald (PME) method was employed [126]. It calculates long range interactions by distributing the charges in the simulation onto a grid which in turn reduces the number of interactions that needs to be calculated. As a result, computational efficiency increases and scales as $NlogN$, as opposed to N^2.

3.3.4 Thermodynamic conditions

The idea of molecular dynamics is to simulate the system at the conditions that are as close as possible to the natural environment. The temperature and pressure, together with volume and number of particles in the system, are controlled using thermodynamic ensembles implemented by various MD integrators.

We can keep the temperature in the system constant by constraining it to a desired value. This is achieved by canonical ensemble (NVT), in which the number of particles (N) and volume of the system (V) are also kept constant. There is also microcanonical ensemble (NVE) in which the energy of the system (E) is fixed, while in isothermal-isobaric ensemble (NPT) pressure (P) is controlled.

There are several methods to regulate temperature in the system [127–129]. In this work, the **Langevin thermostat** was used. It introduces *random force* than puts energy into system and an additional *friction force* which depends on the particle velocity $\mathbf{v}$ and friction coefficient ξ

$$\mathbf{F} = -\xi\mathbf{v} \tag{3.21}$$

The system is behaving as if it is immersed in the heat bath of smaller particles that produce friction on the system. A friction coefficient tunes how quickly the system equilibrates to the desired temperature.

Pressure was regulated using the **Barendsen barostat** [130] which keeps the pressure constant by changing the dimensions of the simulation box. This is achieved through the scaling parameter λ

$$\lambda = (1 + \frac{\delta P}{\tau_P}(P(t) - P_{bath}))^{1/3} \tag{3.22}$$

which depends on relaxation constant τ_P, P(t) pressure at time t, and P_{bath} the pressure of the bath. The pressure is calculated through the position of the particles and forces acting on them, by scaling each of the the atomic directions together (isotropic scaling) or independently (anisotropic scaling).

3.4 Enhanced Sampling Methods

Sampling phase space is a key challenge of molecular dynamics. Phase space is defined as a 6N dimensional space of all positions and momenta in the system. MD studies the time evolution of the system which can be seen as a displacement from one point in phase space to another. However, points in the phase space can be separated by high energy barriers which can be hard to cross with classical MD simulation [131, 132].

Moreover, classical molecular dynamics simulations are performed at physiological temperature and in an explicit solvent model to approximate the natural environment as closely as possible. Under these conditions, the system is sampling phase space for a chosen simulation time, usually in ns - ms range. However, most biological processes happen on a time scale rarely accessible to classical simulations (μs - s) despite the progress in computational power [93]. This has led to the development of different methods that accelerate the sampling in the system we want to study [133].

For example, MD simulations have been used to study conformational changes of biomolecules. However, these processes are characterised by a rugged energy surface where the jumps from one local minima to another are rare on the simulation time scale. This has been referred to as *trapping* and it leaves the large part of the biomolecule phase space unexplored because of the high energy barriers that are hard to cross.

To overcome this issue, many enhanced sampling methods have been developed [132, 134]. Here, a review of several enhanced sampling methods will be given. The list of the reviewed methods has been chosen based on the different approaches used to address sampling issues:

- knowledge of the end state: Umbrella Sampling, Metadynamics
- introduction of the bias potential: Accelerated Molecular Dynamics
- using hybrid MD/Monte Carlo scheme: Replica Exchange

The way the methods are classified is not the only one since there is an overlap between them. For example, Metadynamics requires the knowledge of end state, and the bias potential is constructed during the simulation.

Each of the methods have advantages and disadvantages, and a detailed discussion of these methods is given below.

3.4.1 Umbrella Sampling

The free energy between the states A and B can be obtained by calculating the probability distribution along the reaction coordinate

$$\Delta F = F_B - F_A = k_B T ln \frac{P_A}{P_B} \tag{3.23}$$

where P_A and P_B are the probabilities of finding the system at state A and B, respectively. These probabilities are directly proportional to the time the system spends in each state during an ergodic MD simulation. However, if the two states are separated by a high barrier, a simulation starting in state A is likely to sample only the configuration space around A while sampling of state B is unlikely. Umbrella sampling tries to overcome this issue by introducing a bias potential which is added to the Hamiltonian of the system

$$H = H_0 + \omega(\xi) \tag{3.24}$$

where H_0 is true Hamiltonian and $\omega(\xi)$ is a bias potential along reaction coordinate ξ [135]. By properly choosing the bias potential, the system is forced to sample particular regions of configuration space. The most common choice is a harmonic potential

$$\omega(\xi) = \frac{k}{2}(\xi - \xi_0)^2 \tag{3.25}$$

which is, in practice, used to restrain the reaction coordinate to various values. However, free energy along the reaction coordinate, which is called the Potential of Mean Force (PMF), is now computed not for the system we want, but for the biased system.

Umbrella sampling is able to overcome this issue by reweghting the biased data where the PMF is defined as

$$F^{(u)}(\xi) = F^{(b)}(\xi) - \omega(\xi) + \Delta F \tag{3.26}$$

where ΔF is the free energy of introducing the bias, $F^{(b)}(\xi)$ is the biased and $F^{(u)}(\xi)$ is the unbiased free energy of the system.

To recover unbiased simulation from the biased, the Weighted Histogram Analysis Method (WHAM) [136] is used where the free energy ΔF is calculated by

$$e^{-\beta\Delta F} = \int e^{-\beta\omega(\xi)} P(\xi) d\xi \tag{3.27}$$

In Umbrella sampling several independent simulations are run with different biases. The reaction coordinate is divided into several windows and the bias is calculated for each window. Then, by using WHAM, a histogram is created by calculating a relative probability of observing the states of interest and the free energy ΔF is obtained by using Equation 3.27.

3.4.2 Metadynamics

Another method which requires the knowledge of the end state is called Metadynamics [137]. Here, a history dependent potential fills the free energy minima and the system explores the configurational space. Unlike Umbrella Sampling, it explores the low energy regions first. In Metadynamics the bias potential acts on the chosen collective variables (CV) and is constructed on-the-fly during the simulation. Let ξ be the set of d functions of the microscopic coordinates of the system:

$$\xi(R) = (\xi_1(R), \xi_2(R), ..., \xi_d(R)) \tag{3.28}$$

The Hamiltonian has a form

$$H(q, p; t) = H(q, p) + V(t, \xi) \tag{3.29}$$

where H(q,p) is a Hamiltonian of unbiased system and V(t,ξ) is a bias potential which is in the form of Gaussian function and is added every τ_p steps to flatten the free energy surface (Figure 3.1). The bias potential depends on the parameters defining the height of the added hills and the rate at which they are added - the deposition rate. These parameters define the accuracy and the rate of reconstructing the free energy profile. In the long time limit, Metadynamics assumes that the bias potential converges to the free energy as

$$\lim_{t\to\infty} V(\xi,t) = -F(\xi) + C \tag{3.30}$$

The reconstructed Metadynamics free energy profile compared to the true free energy was proven to have error [138]

$$\epsilon \sim \sqrt{\frac{\omega}{D\beta}} \tag{3.31}$$

where D is diffusion coefficient in the CV space, $\beta = 1/k_B T$ and ω is deposition rate.

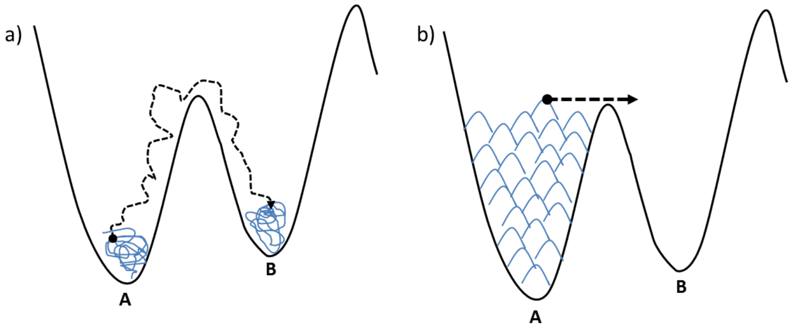


Figure 3.1: Two energy basins A and B are separated by a high energy barrier. a) In normal MD the system is usually stuck in energy minima, while in Metadynamics (b) small Gaussians are constructed on the fly and the system is able to escape the energy minimum from A to B.

As can be seen in the Equation 3.31, error in free energy surface (FES) depends on the deposition rate ω, the ratio of height and frequency at which Gaussians are added, which defines how fast the energy minima are filled with the biasing potential. The problem with this parameter is that we simultaneously want to fill minima quickly (big ω) and decrease the final error (small ω) [139]. This problem was overcomed by introducing another version of Metadynamics called Well-Tempered Metadynamics where the deposition rate decreases over simulation time, $\omega \sim 1/t$ [140]. Moreover, the form of the bias potential was also changed, and it does not converge directly to free energy but with dependence on the temperature

$$\lim_{t\to\infty} V(\xi,t) = -\frac{\Delta T}{T+\Delta T}F(\xi) + C \tag{3.32}$$

where ΔT is the input parameter with temperature dimensions. Therefore in the limit $\Delta T \to 0$ normal MD is recovered while in case $\Delta T \to \infty$ standard Metadynamics is recovered.

To summarise, the efficiency of Metadynamics depends on a good choice of collective variables. If we chose those CVs that do not include the slow motion of the system then the system will not explore the CVs as efficiently as it should [141]. It is non trivial to choose good CVs, and it is based more in experience, chemical intuition or experimentation. Moreover, the error in the reconstructed free energy profile in Metadynamics depends on the number of CV variables used because it scales exponentially with the number of CVs included, so only a small number of dimensions can accurately reproduce the free energy surface.

3.4.3 Accelerated Molecular Dynamics

Compared to other methods, Accelerated Molecular Dynamics (AMD) [142] does not require the knowledge of the underlying free energy surface nor does it require prior choice of a set of reaction coordinates. The idea is to add a bias potential $\Delta V(\mathbf{r})$ to the true potential energy of the system V(r) (Figure 3.2). The extent to which the potential energy surface is modified depends on the difference between the bias potential and true potential. The modified potential $V^*(\mathbf{r})$ is of the form

$$V^*(\mathbf{r}) = \begin{cases} V(\mathbf{r}), & V(\mathbf{r}) \geq E \\ V(\mathbf{r}) + \Delta V(\mathbf{r}), & V(\mathbf{r}) < E. \end{cases} \tag{3.33}$$

where V(r) is potential energy of the system, $\Delta V(\mathbf{r})$ is bias potential, and E is energy threshold. They are schematically shown on the Figure 3.2, and further explained below. AMD allows the boosting of either the whole potential energy of just the dihedral part with equation

$$\Delta V(\mathbf{r}) = \frac{(E - V(\mathbf{r}))^2}{\alpha + (E - V(\mathbf{r}))} \tag{3.34}$$

To define the bias potential, a threshold energy E and acceleration parameter α need to be defined (Equation 3.34). In order to do this, a short MD simulation

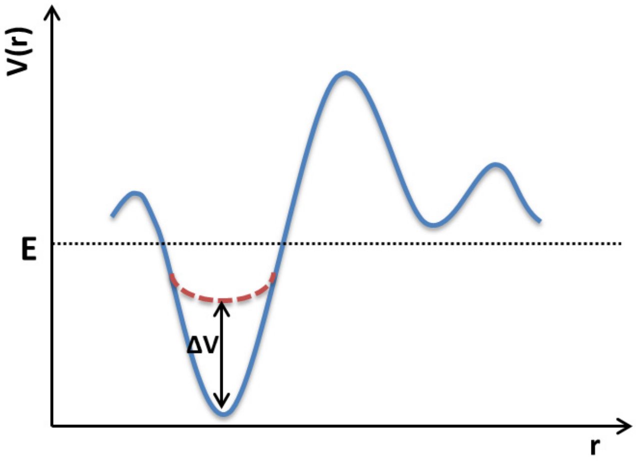


Figure 3.2: Schematic representation of the normal potential V(r) of the system, bias potential ΔV(r) and the threshold energy E.

should be run. From this, the average potential energy $V(\mathbf{r})$ is obtained which is then summed with the approximate energy contribution per degree of freedom. For example, if the system is comprised of 64 residues and 3.5 kcal/mol/residue is the energy contribution per residue, then the energy contribution from all the residues equals 224 kcal/mol. This value is then summed up with the average dihedral or total potential energy of the system obtained from the short MD simulation, and that is the recipe to calculate the threshold energy E while α equals one fifth of this value [143].

For each enhanced sampling method it is important to yield the correct canonical averages of thermodynamic value $A(\mathbf{r})$. In this case, reweighting the biased potential data to extract the underlying unbiased results from the biased trajectory is achieved by multiplying the modified potential by the Boltzmann factor of the bias potential $e^{\beta\Delta V(\mathbf{r})}$.

$$\langle A^C\rangle = \frac{\int d\mathbf{r}A(\mathbf{r})e^{-\beta V(\mathbf{r})-\beta\Delta V(\mathbf{r})}e^{\beta\Delta V(\mathbf{r})}}{\int d\mathbf{r}e^{-\beta V(\mathbf{r})-\beta\Delta V(\mathbf{r})}e^{\beta\Delta V(\mathbf{r})}} \tag{3.35}$$

$$\langle A^C\rangle = \frac{\int d\mathbf{r}A(\mathbf{r})e^{-\beta V(\mathbf{r})}}{\int d\mathbf{r}e^{-\beta V(\mathbf{r})}} = \langle A\rangle \tag{3.36}$$

where the A^C is correct thermodynamic value of variable A(r). The last equation shows that AMD method converges to the canonical distribution after reweighting of the conformational space. The method has been proven to enhance sampling on the systems of different complexity [142, 144–146].

3.4.4 Replica Exchange Methods

Replica Exchange (RE) methods are one of the most popular enhanced sampling methods [134]. They accelerate sampling either by simulating the system across a range of temperatures [147] or by modifying the underlying Hamiltonian [148]. Unlike other methods reviewed here, they do not require the knowledge of the potential energy surface or collective variables *a priori*. Moreover, the result of the simulation is a Boltzmann weighted ensemble and thus no post-processing reweighting is required.

Here, the review of two variants of the Replica Exchange methodology will be given, Temperature RE and Solute Tempering RE. Both are used in this work.

3.4.4.1 Temperature Replica Exchange Molecular Dynamics

The general idea of Replica Exchange Molecular Dynamics (REMD) [147] is to simulate a series of independent replicas of the original system across a range of different temperatures, usually between 250 K to 600 K. At high temperatures, the system has higher kinetic energy, and therefore it is able to sample larger volume of the phase space, while at lower temperatures the system may become trapped in local minima. However, REMD overcomes the problem of bad sampling at lower temperatures by allowing the systems at different temperatures to exchange their configurations. Since we are interested in the simulation results in the range of physiological temperatures, replicas simulated at these temperatures contain the configurations from the whole temperature space thus ensuring that the system has sampled more phase space than it would with normal MD [149]. However, simulation of N replicas, compared to one in normal MD simulation, requires more computational power. Therefore REMD is limited by computational expense as it requires access to a highly parallelised supercomputer.

Theory. In this method, the system is comprised of N non-interacting replicas (copies) at M different temperatures. A state in REMD is described as

$$X = x_m^{[i]} = (\mathbf{r}^{[\mathbf{i}]}, \mathbf{v}^{[\mathbf{i}]})_m \tag{3.37}$$

where subscript m and superscript in square brackets [i] label the replica and temperature, respectively. In the canonical ensemble, the probability of a state existing at a given temperature, $W(x)$, is weighted by Boltzmann factor as shown in Equation 3.37. A β is inverse temperature and Q is the partition function - sum of all states.

$$W(x) = \frac{e^{-\beta E(\mathbf{r},\mathbf{v}})}{Q} \tag{3.38}$$

Similarly the probability of state X in the *generalized ensemble* can be expressed as the product of the Boltzmann factors for each replica.

$$W_{REMD}(X) = \frac{e^{-\beta_1 E_1 - \beta_2 E_2 \ldots \beta_s E_s}}{Q_{REMD}} \tag{3.39}$$

In order for the exchange process to converge to equilibrium, the *detailed balance condition* is imposed. It says that the chance of exchanging between two states must be identical; the probability of making the move, times the probability of accepting the move, must be identical in the forward and reverse directions

$$W_{REMD}(X) w(X \rightarrow X^{'}) = W_{REMD}(X^{'}) w(X^{'} \rightarrow X) \tag{3.40}$$

where w is the probability of accepting the move, and W is the probability of being in the state, and the chance of attempting the move is identical in the two directions.

Furthermore, a Monte Carlo Metropolis test is derived to swap replicas based on their potential energies and temperatures (Equation 3.41). Since the acceptance probability decreases exponentially with temperature difference, only neighbouring replicas are exchanged

$$w(X \rightarrow X^{'}) = \begin{cases} 1 & if \Delta \leq 0; \\ exp(-\Delta) & if \Delta > 0; \end{cases} \tag{3.41}$$

where
$\Delta \equiv [\beta_n - \beta_m](E_p(r^{[j]}) - E_p(r^{[i]}))$

Prior to a REMD simulation, each replica is equilibrated to a chosen temperature. Each replica is then run for a defined time (time between exchanges, usually 2 ps) in the NVT ensemble, and adjacent replicas are swapped if the Metropolis criterion is satisfied. By repeating this process, movement in temperature space is achieved and the potential energy surface is explored. If a replica approaches an energy barrier, its potential energy will increase and it is likely to swap to a higher temperature. Moreover if at this higher temperature the replica is able to overcome the energy barrier it is then likely to swap back to lower temperature.

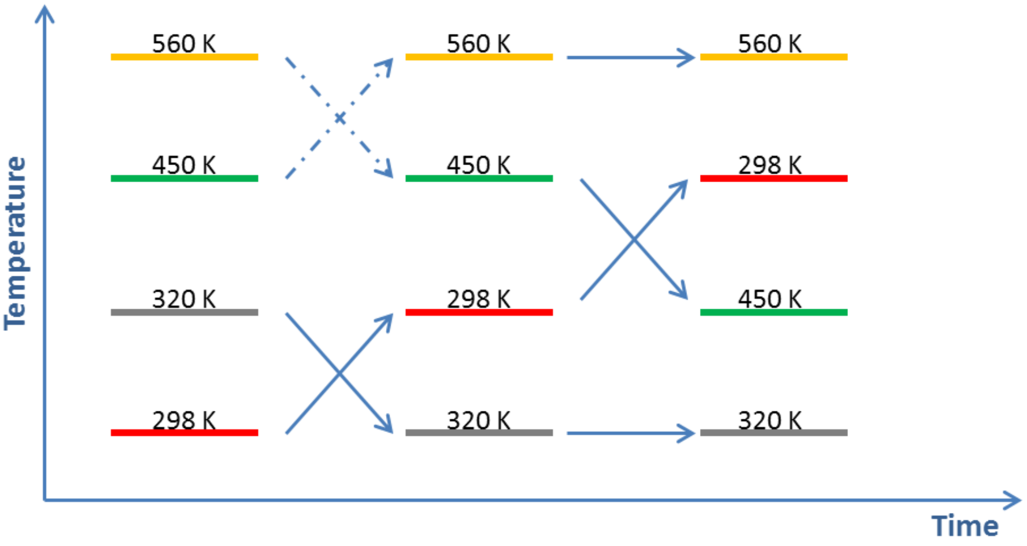


Figure 3.3: Example of the REMD simulation. A dotted arrow indicates a failed test, while a solid arrow indicates a successful move.

Therefore a random walk in the conformational space is induced by swapping the replicas.

A typical REMD trajectory is shown in Figure 3.3. This testcase is composed of four different replicas starting at four different temperatures, 298 K, 320 K, 450 K and 560 K. Each block represents a normal MD simulation which is run for certain time after which the Monte Carlo test is attempted and replicas have either swapped or not based on their potential energies and temperatures. For example, a replica at T = 298 K passes two tests, moving upward in temperature each time. On the other hand, a replica starting at T = 450 K did not pass the test in the first instance so it continued to run at the same temperature while at the second test it moved downward swapping places with the replica at T = 450 K. The replicas at the minimum and maximum temperature are tested half as often as others, as the test is applied only to adjacent replicas.

However, proper performance of REMD depends on the several points [147]:

1. Are the temperatures optimally distributed?

2. Is the number of replicas (temperatures) sufficient?

3. Is the highest temperature high enough to pass the high energy barrier?

An important factor affecting the efficiency of the REMD algorithm is the temperature distribution of replicas. If the temperatures are closely spaced then replicas will exchange frequently due to the high overlaps of their potential energy distributions, but many replicas will be required to span a given temperature interval. If spaced far apart, then fewer replicas are required, but the acceptance probability will be reduced. For maximum efficiency, a uniform acceptance profile is desired. Therefore, the acceptance probability is a compromise between the method expense and rate of convergence. A probability of acceptance 0.2 - 0.4 was shown to be sufficient to ensure good mobility of replicas [150].

The number of replicas is also affected by the system size since the number of required replicas scales in order of $\sqrt{D}$, where D is the number of degrees of freedom in a system [151]. The system size is mostly influenced by the type of solvent used, especially if the explicit solvent is chosen as was in our case. Then the number of replicas ranges between 50 and 100, or even more depending how high a temperature we want to simulate will be, which has important influence on the simulation time necessary to get converged results. Furthermore, it is not trivial to check if the highest temperature is high enough to pass the high energy barrier. It mainly depends on the system of interest.

3.4.4.2 Replica Exchange with Solute Tempering

Although the temperature Replica Exchange is the most commonly used tempering method because of its ease to use and implementation in all major molecular dynamics software packages, other replica properties can also be used to enhance the sampling of the system. In particular, to facilitate the sampling over the rugged energy landscape, the specific interactions within the system can be softened, such as scaling the strength of hydrogen bonds or hydrophobic interactions [151]. In this way, sampling is performed on the smoothed energy landscape, and compared to other enhanced sampling methods, only few system degrees of freedom are softened.

The Replica Exchange with Solute Tempering (REST) method is an example of Hamiltonian replica exchange method where the Hamiltonian of the system is scaled down to obtain sufficient sampling of the system [148]. In REMD the whole system is simulated across given temperatures, and the number of replicas is proportional to the number of the degrees of freedom. On the other hand, by modifying the Hamiltonian as is done in the REST method, it is possible to enhance sampling of only parts of the system in which we are interested, keeping the rest of the system (usually water molecules) at room temperature.

REST has approached this problem by dividing the system into two parts. It is usually done in such a way that the molecule of interest (usually the whole solute, although only parts of it could be used) is assigned as “hot” and is subject to the scaling of the Hamiltonian, while the rest of the system (usually solvent) is kept at the “cold” temperature. This can be accomplished because the Hamiltonian is an additive property and can be decomposed into energy terms contributing to the total energy of the system. In REST, the total potential energy of the system is composed of the three parts; the energy of the solute or central part (E_p), the interaction energy between the solute and the solvent (E_{pw}), and the interaction energy between the solvent molecules (E_{ww}).

$$E_0(X) = E_p(X) + E_{pw}(X) + E_{ww}(X) \tag{3.42}$$

where X is the configuration of the whole system. The potential energy of the system for the replica m is scaled as

$$E_0(X) = \left[\frac{\beta_m}{\beta_0}\right] E_p(X) + \sqrt{\frac{\beta_m}{\beta_0}} E_{pw}(X) + E_{ww}(X) \tag{3.43}$$

where $\beta_m = 1/k_B T_m$, $\beta_0 = 1/k_B T_0$, T_0 is the lowest temperature, while T_m is the temperature of the m-th replica.

A scaling factor $\lambda = \dfrac{\beta_m}{\beta_0}$ is used to scale the interactions in the system [152]. The Hamiltonian of the solute atoms is parametrised as following:

- The charge of the atoms is scaled by a factor $\sqrt{\lambda}$
- The Lennard Jones parameter ϵ is scaled with λ
- The force constants of the bonded terms are scaled by λ

In this way, the scaling is achieved by multiplying the solute intramolecular potential energy by a lambda factor in order to lower the energy barriers [152]. The scaling factor adopts values between 0 and 1 (unscaled potential for the lowest temperature). The first two terms in the equation 3.43 are made of only small number of degrees of freedom and are the reason why fewer replicas are needed to run the system compared to REMD, where the whole system degrees of freedom are utilised.

Just like in REMD, the replicas are obeying Metropolis criterion and detailed balance condition. However, since only the lowest replica is not run on the modified potential energy surface, it is the only one from which the canonical ensemble can be obtained, while others are used to facilitate the sampling.

To get a number of replicas, the geometric distribution based on a scaling factor $\lambda = \dfrac{\beta_m}{\beta_0}$ is used. The geometric distribution of the replicas is calculated as

$$T_m = T_{min} * exp^{m * \dfrac{log(T_{max}/T_{min})}{(m-1)}} \tag{3.44}$$

where T_m is the temperature of the m-th replica, T_{min} is the minimal temperature used, T_{max} is the maximum temperature used.

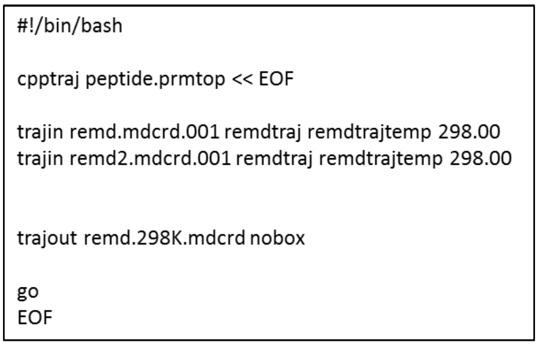
The main advantage of the REST over REMD is a lower number of replicas required to efficiently sample the system, which makes it more attractive to study larger system like membranes [153] or smaller challenging systems such as intrinsically disordered peptides [154]. Overall, smaller number of replicas thus give a way to a shorter running time on a supercomputer which results in faster real time to get simulations done than it is the case with REMD.

3.4.5 REMD vs. REST post processing

REMD and REST are the two enhanced sampling methods used in this work. However, they are implemented in different MD packages, so their post processing is different. REMD used in this work was run in Amber, while REST was run in Gromacs with the Plumed patch [155]. In both methods, we were interested at extracting a constant temperature trajectory from T=298 K.

The method implementation differs in such a way that at the exchange attempt, an Amber exchanges temperatures while Gromacs exchanges coordinates. Because of this, Amber simulation requires post processing using Amber module *cpptraj* which is building a constant temperature trajectory from the frames belonging to the specified temperature contained in different replica trajectories. The idea of the temperature exchanges is shown in Figure 3.3 where it can be seen how temperature replicas are travelling up and down the replica ladder, while we are only interested to extract a constant temperature trajectory at T=298 K.

The way Amber post processes REMD simulation with the *cpptraj* module is using the *trajin* command which takes the first replica trajectory from all simulation repeats along with the command *remdtrajtemp* specifying the temperature, and then extracts the trajectory from all the simulations. The example file is shown on the Figure 3.4.



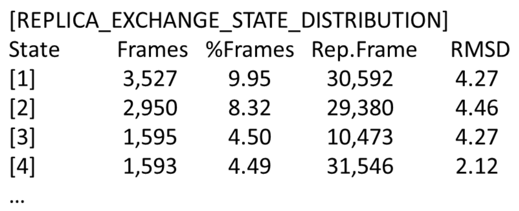
```
#!/bin/bash

cpptraj peptide.prmtop << EOF

trajin remd.mdcrd.001 remdtraj remdtrajtemp 298.00
trajin remd2.mdcrd.001 remdtraj remdtrajtemp 298.00


trajout remd.298K.mdcrd nobox

go
EOF
```

Figure 3.4: The example of the *bash* script used to analyse REMD Amber simulations. Here the example is given for two simulation repeats.

Gromacs, on the other hand, exchanges coordinates instead of temperatures, so the ensemble at the temperature of interest (T=298 K in our case), is usually the lowest replica trajectory. In Gromacs, using the *trjcat* command, all the lowest replica trajectories are concatenated to get a final temperature trajectory at T=298 K which was subsequently used for analysis.

3.5 Analysis methods

In this section, the background behind a few methods used to analyse the simulation data will be given. These include the Dash software used to analyse cluster conformations based on torsion angle values, the Dashsim program, which using circular similarity, compares torsion values to each other, and a few statistical metrics used to assign similarity between data sets.



```
[REPLICA_EXCHANGE_STATE_DISTRIBUTION]
State     Frames  %Frames  Rep.Frame   RMSD
[1]        3,527    9.95    30,592     4.27
[2]        2,950    8.32    29,380     4.46
[3]        1,595    4.50    10,473     4.27
[4]        1,593    4.49    31,546     2.12
...
```

Figure 3.5: Example of the part of Dash output file where each Dash state (first column) has associated number of frames, the total population of these frames in the trajectory, the representative frame in the trajectory and the RMSD value.

3.5.1 Dash software

In this work, the main tool used to analyse the simulations was the Dash software which performs torsion based clustering [156]. The torsions of interest are extracted from the trajectory and saved in an input file as a *torsion time series*. The file is then run with the Dash software, which gives as the output a file that contains:

- A list of Dash states obtained by clustering the torsion space
- Population of each Dash state in the trajectory
- The main torsion values and associated standard deviations for each state
- A representative structure for each state
- The Dash states time series

A Dash state is a torsion angle ensemble characterising a distinct conformation occurring once or several times for a certain amount of time during the simulation. Each Dash state is characterised by the mean and SD of the torsion values, so it is then possible to compare the states due to their similar torsion values to get the final list of unique Dash states. Moreover, with each Dash state is associated a frame from the trajectory as a representative structure (Figure 3.5), which makes it possible to visualise and double check if the states that are assigned as similar are truly belonging to the same cluster, which was always true. Another advantage of Dash includes a list of Dash states time series, which was used in our analysis to see the time evolution of the different cluster states during the simulation. All these points proved to be satisfactory for our system, so it was decided to use Dash as the main tool in the cluster analysis of our simulation data.

In the software, the parameter called *bout length* defines the minimum lifetime of the torsion angle to make one state. It is defined as

$$\text{state lifetime} = \frac{\text{bout length } (l)}{\text{time step (n)}} * \text{simulation time} \tag{3.45}$$

As Dash was initially intended for MD trajectory where the exchanges between the states are rare due to sampling problem, the bout length was defined to that kind of trajectory. However, the definition was later adjusted to the REMD or REST trajectory where due to the nature of the exchanges, the lifetime of the state between the exchanges can vary. The procedure of finding Dash states is thus reduced to:

- Find the (micro) states for the individual torsion angles
- Combine the torsion angle states into states for the whole system
- Calculate the number of frames spent in each combined state. By including special flag before running Dash command, the states populated less than 1% of the total time will be included in the final list of states to account for the REMD or REST nature of exchanges.

In total, there are two versions of Dash software made to analyse MD and REMD or REST trajectories, and both versions are freely available to be downloaded from the University of Portsmouth webpage.

3.5.2 Circular statistics

In general, the mathematical data can be treated as either linear or circular (directional). The need for circular statistics has arisen because of the quantities like torsion angles or daytime where the simple arithmetic mean is not appropriate to use. For example, the mean of the angles 0° and 360° is not 180° because 0° and 360° are the same angles on the unit circle.

The way the circular statistics treats such data is that it converts polar coordinates to Cartesian and then the mean of an array of angles θ_i is calculated as

$$Circular\ Mean = arctan(S/C) \tag{3.46}$$

where

$$S = \sum sin(\theta_i) \tag{3.47}$$

$$C = \sum cos(\theta_i) \tag{3.48}$$

The standard deviation is defined as

$$Circular\ Standard\ Deviation = \sqrt{-2log(R)} \tag{3.49}$$

where

$$R = \sqrt{(S^2 + C^2)} \tag{3.50}$$

In this work, circular statistics was used to calculate the mean and SD of a set of torsion angles.

The circular statistics is also a foundation of the program **Dashsim** that was used to compare similar torsion angles. A comparison is possible if the number of torsions is identical and the Dash output format is used. The Dashsim program calculates the similarity matrix between two sets of Dash states. The output contains a matrix of values that lie in [0, 1] range where 1 means that two states are completely the same. In the cluster analysis that was performed on our simulation data, two sets of torsions were considered to be similar and assigned to the same cluster conformation if the circular similarity between them was ≥ 0.65. A Dashsim script is written by David Whitley from the University of Portsmouth.

3.5.3 Statistical measures of similarity

In this work, three commonly used statistical measures of similarity were used to test the similarity between the data sets. These are Mean Signed Error (MSE), Mean Unsigned Error (MSE) and coefficient of determination (R^2).

The similarity metrics were used to examine the differences between the experimentally measured and theoretically calculated chemical shifts.

Linear regression. The simplest way to get the correspondence between two data sets is to plot one against each other, which is mathematically expressed as linear regression

$$y = kx + b \tag{3.51}$$

where k is a slope, and b intersection. The best fit line through the data points is called regression line. If data points are very close to the regression line on the plot, it means that data pairs have values that are close to each other. The parameter which describes such fit is called **coefficient of determination** (R^2). It takes values between 0 and 1, where 1 means that data match perfectly.

Mean Signed Error (MSE) calculates the mean of the difference between the pairs in the entire dataset. It tells us how far up or down from the average the data is.

If we have two data sets, one set of experimental values, and the other one computational values, the difference between all the data pairs is calculated (Equation 3.52), and then averaged over entire set gives MUE value (Equation 3.53)

$$\Delta\delta = \delta_{calc} - \delta_{exp} \tag{3.52}$$

$$MUE = \frac{\sum_{n=1}^{N} \Delta\delta_n}{N} \tag{3.53}$$

where $\Delta\delta$ is the data pairs difference, and N total number of data pairs. We can also get the standard error of the mean of the MUE as

$$SE = \frac{SD}{\sqrt(N)} \tag{3.54}$$

where SD is a standard deviation of the sample, and N is total number of data pairs as before. The standard deviation is the square root of the average of the squared deviations from the mean

$$SD = \sqrt{\frac{\sum_{n=1}^{N} (\Delta\delta_n - \overline{\Delta\delta})^2}{N-1}} \tag{3.55}$$

where the $\overline{\Delta\delta}$ is the mean of the data pairs difference.

Mean Unsigned Error (MUE) is similar to MSE, but differs only that it calculates the absolute difference between data pairs $\Delta\delta$.

$$MUE = \frac{\sum_{n=1}^{N} \mid \Delta\delta_n \mid}{N} \tag{3.56}$$

The MSE is more significant than MUE in a way that it tells how far away the data pairs are from the mean. The standard error of the mean of MSE can also be calculated using the Equation 3.54.

Bootstrapping [157] is a method used to estimate any property of the sample by measuring the observables from the limited or approximate distribution.

The method is performed in such a way that a *bootstrap sample* is obtained by randomly sampling n times, with replacement, the values from the original data. Then, the bootstrap algorithm generates a large number of independent bootstrap samples, each of size n. It is usually generated 10,000 (10k) bootstrap samples, and for each of the bootstrap samples, the value of the statistic we are interested into is calculated, for example the mean of the data. Then, the estimate of the standard error is calculated for the bootstrap statistic using the standard deviation of the bootstrapped resampling distribution.

An estimate of the 95 % confidence interval (CI) of the bootstrapped statistic metrics is calculated as

$$\overline{\theta} \pm 1.96 * SE \tag{3.57}$$

The SE is multiplied by 1.96 to obtain an estimate of where 95 % of the population sample means are expected to fall in the normal distribution.

Here the general idea of the bootstrapping was described. Each statistical method explained previously, MUE, MSE and R^2 were bootstrapped with 95 % CI to examine how well computational chemical shifts fit the experimental data. More detailed discussion of the bootstrapping applied in this work is given in Chapter 4.7.2.

3.6 Summary

In this chapter, a background of the computational methods was given. First the QM theory was introduced, followed by the review of the molecular dynamics and enhanced sampling methods. In this work, the conformational ensemble of the cyclic peptides was explored using two enhanced sampling methods, temperature Replica Exchange and Solute Tempering. Then, QM based DFT approach was used to calculate chemical shifts on the representative peptide structures that were then compared with the experimentally measured chemical shifts using the statistical metrics described in section 3.5.3. Therefore, a combination of results from enhanced sampling methods, DFT calculations and NMR experiments was used to obtain the conformational ensemble of the studied peptides. In the next few chapters, the results obtained for the each studied peptide using introduced methodologies will be described.

Chapter 4

8-Arg Vasopressin

8-Arg Vasopressin, usually referred to as AVP, is a 9 amino acid long cyclic peptide composed of cyclic moiety ($Cys^1 - Tyr^2 - Phe^3 - Gln^4 - Asn^5 - Cys^6$) connected by disulphide bridge ($Cys^1 - Cys^6$), and a tail ($Pro^7 - Arg^8 - Gly^9$) capped with the NH_2 group [158]. The C terminal NH_2 group is a natural form of AVP, also visible in the X-ray AVP structure [159]. The total +2 charge of AVP comes from the N terminal cysteine together with the guanidinium group of Arg^8 (Figure 4.1).

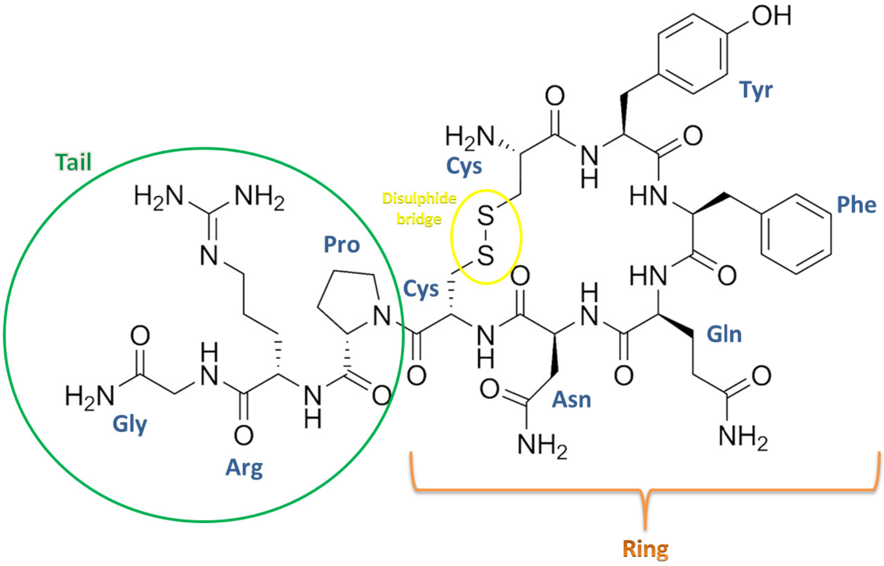


Figure 4.1: The structure of 8-Arg Vasopressin. Cyclic part is connected with disulphide bridge between the cystein residues, and tail is circled in green.

4.1 Biological background

AVP is a *hormone peptide* released by the posterior pituitary gland as a complex with neurophysin (NP) and secreted into the bloodstream. Vasopressin and Oxytocin are the only two hormone peptides secreted by the pituitary gland that act at a distance. AVP is also known as an antidiuretic hormone (ADH), as it regulates blood osmolarity by keeping the concentration of water, salts and glucose inside the physiological boundaries [160] (Figure 4.2).

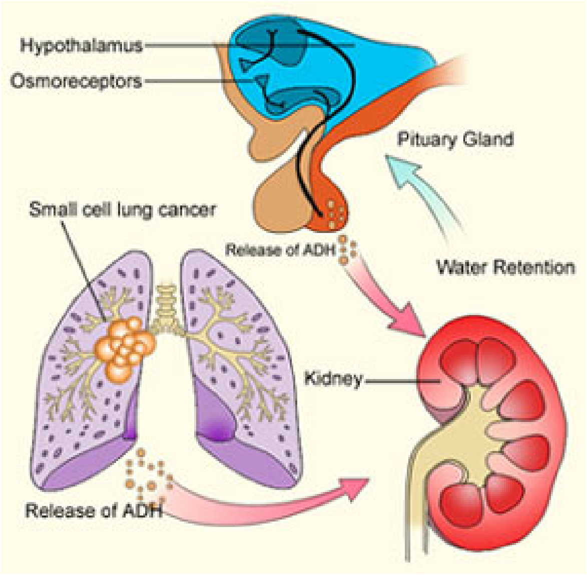


Figure 4.2: The picture shows the mechanism in which AVP acts as an antidiuretic hormone (ADH) [161].

After it is released from the pituitary gland, it binds to the G protein-coupled receptor (V2R) within the kidneys that promotes insertion of aquaporins into the plasma membrane of the kidney collecting duct, where it stimulates water reabsorbtion. The biochemical path involves AVP binding to V2R which stimulates the synthesis of cAMP protein that activates protein kinase A (PKA) resulting in the opening of the aquaporins on the cell membrane. The aquaporins or "water channels" transport solute-free water through tubular cells back into blood, leading to a decrease in plasma osmolarity and an increase osmolarity of urine [160]. However when the kidneys cannot concentrate urine normally, a large amount of dilute urine is excreted which is an indicator of the disease called diabetes insipidus [162]. Hyper production of urine leads to dehydration resulting in increased thirst and a desire to drink. The disease is controlled by using a medicine that resembles

ADH [163]. Therefore, the water balance in our body is regulated by a combination of mechanisms that include AVP secretion, thirst and renal function.

Beside its primary role as an antidiuretic, AVP is also involved in the regulation of blood pressure [164], and it is thought that it mediates social and sexual behaviour, especially aggression, anxiety and pair-bonding [165].

4.2 Experimental studies on AVP

AVP is a well studied peptide whose conformation has been revealed by two crystallographic and several NMR studies. Here an overview of the known experimental data will be given. The focus of all experimental studies has been on revealing the conformation of the ring part of the structure, while the tail part has been characterised as rather flexible.

AVP was fully crystallized as a part of a trypsin complex (PDB:1YF4) [159]. The conformation of AVP in this complex was characterized by an almost planar ring arrangement with no significant hydrogen bonds between the ring residues, and an extended tail (Figure 4.3 (a)).

In another study, the ring part of the neuropeptide 8-Lys-AVP (PDB:1JK4) [166] was crystallized, which shares the ring sequence with AVP but differs in the *Lys-8-Arg* tail mutation. However, the tail was not crystallised in this study. Here the ring conformation was characterized by more *saddle*-like structure compared to the structure crystallised with the trypsin complex, and has β-turns centred at residues 3,4 and 4,5 (Figure 4.3 (b)). This was an X-ray resolved structure in complex with neurophysin (NP), the AVP carrier protein.

A few NMR studies also report on different ring arrangements. They suggest the existence of *saddle*-like AVP ring conformations irrespective of the polarity of solution. The structure in both water [167] and DMSO [168] was characterised by 2 β-turns at positions 3,4 and 4,5. No significant intramolecular hydrogen bonds were found in water, while in DMSO only one hydrogen bond was observed between $Tyr^2O - Phe^5H$. Studies in SDS micelles suggest that the ring conformation of the AVP attached to the micelles appears to be similar to the ring 8-Lys-AVP-NP complex *saddle*-like form with only one intramolecular hydrogen bond observed

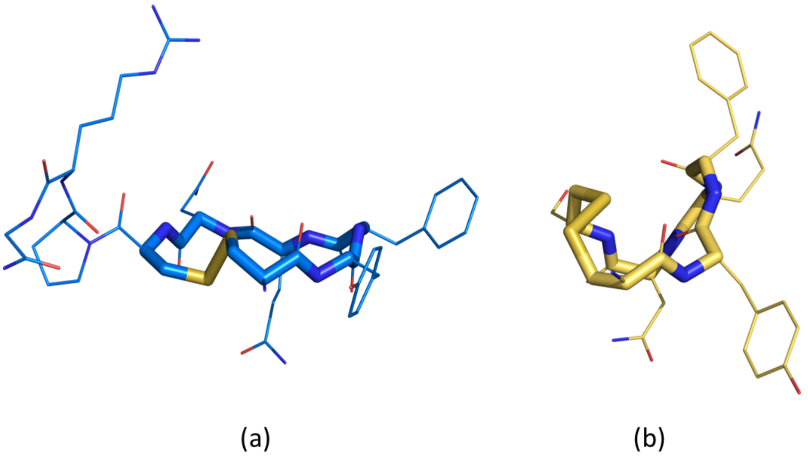


Figure 4.3: (a) The structure of AVP co-crystallised in a trypsin complex (PDB: 1YF4). The ring part of the structure is characterised with almost planar atom arrangement. (b) Crystal structure of the ring part of the 8-Lys-AVP peptide co-crystallised in a complex with neuropysin (PDB:1JK4). The ring structure is characterised as a saddle-like conformation.

between $Phe^3O - Cys^6H$ [169].

Another interesting observation in this study was the behaviour of hydrophilic Arg^8 which shows a tendency to be turned toward the polar environment, promoting extended tail conformations. This observation is important because the interaction between Arg^8 and an extracellular loop of receptor is thought to be a key to receptor recognition [170].

A more recent NMR study in DPC micelles suggested a ring conformation with β-turns at residues 3,4 and 4,5 with a $Tyr^2O - Cys^6H$ intracyclic hydrogen bond. The tail was described as creating a 6,7 β-turn for almost 3/4 of the observed conformations [171].

To summarise, the experimental studies suggest that AVP adopts flexible backbone conformations which probably helps it to perform different biological functions. Two types of ring conformations were observed - planar and saddle-like, while the tail can be elongated or folded toward the ring. In the ring saddle-like conformation, the presence of a few hydrogen bonds was observed, Tyr^2O-Phe^5H, $Tyr^2O - Cys^6H$ and $Phe^3O - Cys^6H$.

4.3 Computational studies on AVP

Besides being studied by experimental techniques, the conformation of AVP was also studied with computational methods [172–174]. The first of these methods was done in 1996 [172], when a combination of Monte Carlo method and molecular dynamics was used to check for the conformational flexibility of AVP and Oxytocin. The simulation suggested conformations of AVP have β-turns centred at residues 3,4 and 4,5, and at residues 2,3 and 3,4 for Oxytocin.

Next study was done with reservoir REMD in ff99SB-ILDN force field and TIP4P-Ew water model, analysed at 298 K [174]. This simulation method differs from temperature replica exchange in such a way that reservoir structures are generated in advance through normal MD simulation at high temperature, and then the configurations are exchanged between reservoir replicas and normal replicas [175]. Using this method, the *canonical* ring structure stabilised by $Tyr^2O - Asn^5H$ and $Tyr^2O - Cys^6H$ hydrogen bonds was observed. Two additional hydrogen bonds were also reported, one between ring residue and tail $Cys^6O - Gly^9H$, and another one between tail residues $Pro^7O - Gly^9H$.

Another study is from Haensele et al., who performed an 11 μs long MD simulation of AVP in explicit water at a temperature of 298 K using the Amber ff99SB force and TIP4P-Ew water model [173]. This MD simulation, starting with the known PDB conformation (PDB: 1YF4), which is referred to as an open conformation here due to the planar ring arrangement, revealed a few apparent changes in RMSD of the AVP which led to the identification of a four distinct ring conformations followed by the fast movement of the tail region [173], Figure 4.4.

The conformations are clustered into groups depending on the structural characteristics of the cyclic part of the peptide:

- The *Open* conformation is a crystallographic conformation (PDB: 1YF4). The ring structure is rather planar compared to other conformations and hence the name. It is not characterised by any intramolecular hydrogen bond or β-turn.

- The *Saddle* conformation matches the resolved 8-Lys-AVP-NP ring structure (PDB: 1JK4). This conformation is characterised by two β-turns type I centred at 3,4 and 4,5, and is stabilised by Tyr^2O - Asn^5H and Tyr^2O -

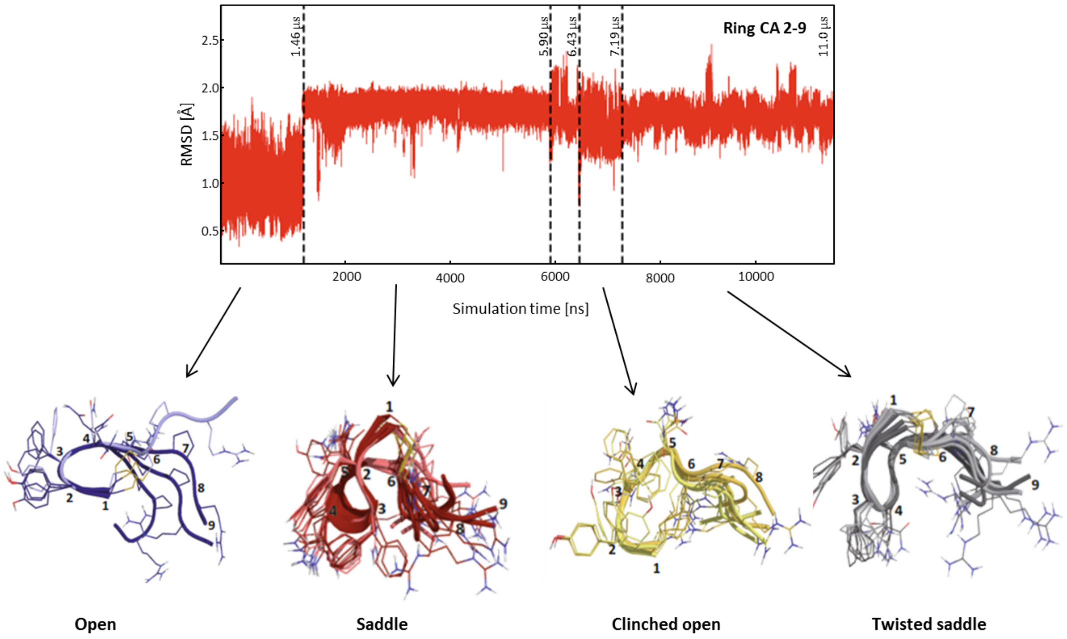


Figure 4.4: Root mean square deviation (RMSD) of 8-Arginine-Vasopressin (AVP) during 11 μs MD simulation in water [173].

Cys^6H hydrogen bonds.

- The *Clinched Open* conformation was characterised with the Phe^3O - Cys^6H hydrogen bond, and a turn centred at residues Gln^4O and Asn^5H.

- The *Twisted Saddle* conformation is mainly stabilised by a hydrogen bond between Tyr^2O and Asn^5H, and a type II β-turn is enclosed by residues Phe^3O and Gln^4H.

The simulation also showed that the tail moves independently of the ring. It exists in two conformations - extended and folded. The extended conformation, characterised with 7,8 β-turn II, appeared 81 % of the total simulation time [173].

Summary of the computational results. The computational studies are in agreement with experimental studies in terms of AVP conformational diversity. AVP is thought to form β-turns at residues 3,4 and 4,5, or not form β-turn at all. The ring conformation is stabilised by $Tyr^2O - Phe^5H$, $Tyr^2O - Cys^6H$ and $Phe^3O - Cys^6H$ hydrogen bonds. The tail part is either folded toward the ring with a 7,8 β-turn and creating $Cys^6O - Gly^9H$ hydrogen bond, or in an extended conformation stretching away from the ring.

Haensele et al. associated AVP ring conformations with the β-turn types, and named the conformations as *Saddle* (3,4 and 4,5 β-turns), *Open* (no turn), *Clinched Open* (4,5 β-turn) and *Twisted Saddle* (3,4 β-turn). They also published the mean values with associated standard deviations of the peptide $\phi\psi$ torsions for the each cluster state, which were later used in this work to compare with the values from our simulations.

4.4 Motivation for our work

The literature review of the AVP data suggests that there are several conformations distinguished in the ring part of the structure, with the tail moving independently of the ring conformations. Interestingly, only one X-ray structure and unrestrained MD simulation suggested planar ring conformation, while this conformation was not reported in any of the NMR data. On the other hand, all the studies report the *saddle*-like ring conformation.

Reviewing the computational methods used, it can be noticed that reservoir REMD reports only *saddle* conformation as well. Another computational method, long timescale MD simulation (11 μs) reported on four main AVP cluster states. However, it was observed that the simulation was not long enough to get converged simulation data regarding the AVP conformational dynamics. The interconversion between the states was very slow, and as such it could not be claimed that the population of the states equilibrated or that the entire phase space of the peptides was explored.

Our goal was then to tackle the problem of the incomplete sampling of the AVP conformational ensemble by using enhanced sampling methods to test for simulation convergence. To achieve this goal, Replica Exchange MD method was chosen, as one of the widely used enhanced sampling methods. The idea was to get converged cluster populations, later to be used to validate against experimental data.

4.5 REMD simulation details

Four sets of simulations were performed using the REMD method with the PMEMD module of the AMBER 12 suite program [176]. Each simulation was started with different starting conformation named as *Open, Saddle, Clinched Open* and *Twisted*

Saddle. The *Open* structure is AVP crystal structures (PDB: 1YF4), while *Saddle*, *Clinched Open* and *Twisted Saddle* initial pdb structures were taken from the previously published MD simulation [173].

The REMD simulations were performed using the Amber ff99SB force field with TIP3P water model [177]. The simulations were prepared by the tleap suite of Amber program where the system was neutralised by adding 2 Cl^- atoms. The Particle Mesh Ewald [126] was used for the long-range interactions using a 10 Å cutoff. Bonds involving hydrogen were constrained using the SHAKE algorithm [121] with a tolerance of 0.00001 Å. REMD simulations were performed in the NVT ensemble using a Langevin thermostat for the temperature coupling with a collision frequency of 1 ps^{-1}. 200 ps of NVT simulation was used to equilibrate the initial state to the desired temperature for each replica. The REMD exchanges were attempted every Δt=2 ps.

The temperature distribution was obtained using an online temperature generator http://folding.bmc.uu.se/remd/ [178]. There, the minimum and maximum temperatures, acceptance probability and the number of atoms in system had to be given. Input was as follows:

- Minimum temperature = 298 K
- Maximum temperature = 550 K
- Acceptance probability = 40 %
- Number of protein atoms = 142
- Number of water molecules = 2409 for *Clinched Open*, 2208 for *Twisted Saddle*, 2351 for *Open*, 2268 for *Saddle*

4.5.1 REMD efficiency

The efficiency of a REMD simulation depends on the capability of the replicas to exchange between lower and higher temperatures. An optimal distribution of temperature induces a free random walk in temperature space. In this work, 80 replicas were used to perform REMD. This number of replicas, together with temperature values, was obtained from the temperature distribution generator to satisfy the requested acceptance probability of 40 %.

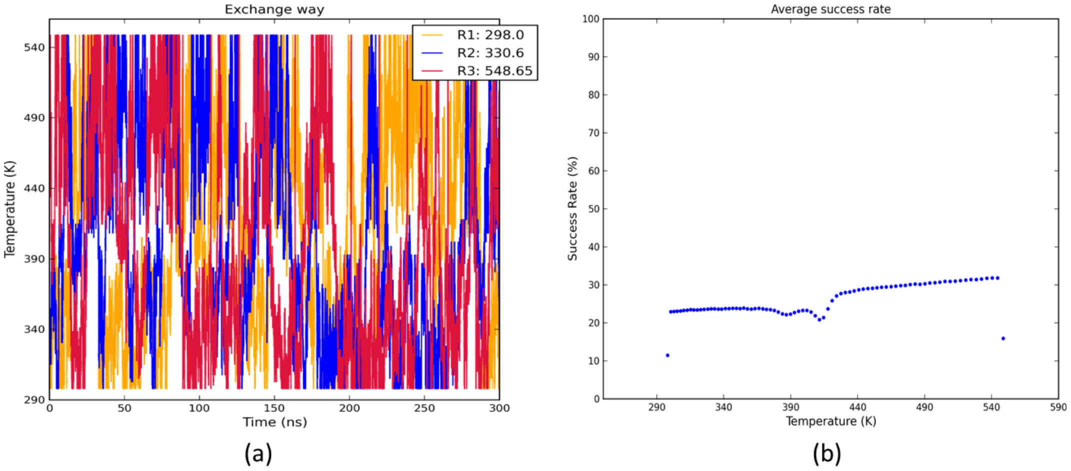


Figure 4.5: (a) Random walk of the three temperature replica (298 K, 330.60 K and 548.65 K) from REMD *Clinched Open* simulation. (b) Acceptance probability between neighbouring replicas.

The observed acceptance probability profile is presented in Figure 4.5 (b). Despite a mild well at around 400 K, an average acceptance probability of 30 % was reproduced across the simulation. Another measure of simulation success is the plot of replica mobility (Figure 4.5 (a)). The data from the REMD *Clinched Open* starting conformation at three different temperatures was chosen to check the mobility: 298 K, 330.60 K and 548.65 K. All temperature trajectories have visited both top and bottom temperatures although there seems to be a boundary at around 400 K where the replicas are stuck either above or under this level. It might be that the system is in a high energy state at this point which cannot exchange down to lower temperatures. This is in agreement with acceptance rate plot where the unexpected minimum also appeared at around T = 400 K.

4.6 REMD simulation results

Four sets of REMD simulations were performed to test for AVP conformational convergence. Our goal was to show that all simulations should give the same result, irrespective of the starting conformation. The property that an identical population pattern is obtained in simulations, independent of starting conformation, is referred to as **convergence** in our case, and that is what was tested with REMD.

The simulations were performed for 300 ns in total per replica. However, an initial 100 ns of each simulation was taken as equilibration time and was not included in the final analysis. The peptides were analysed in terms of β-turn population, hydrogen bond population and cluster state diversity.

The experimental and computational review of the known AVP conformational data revealed that distinct peptide conformations are commonly characterised in terms of β-turn and hydrogen bond populations. On top of that, in the work performed by Haensele et al., the torsion angle based clustering software Dash was used to determine the populations of distinct ring states (Section 4.3). This clustering approach seem particularly relevant in elucidating conformation of the cyclic peptides, so the same method was used in this work too.

A β-turn with hydrogen bond populations, and a cluster state analysis with Dash software are two independent, but complementary analysis methods in the sense that both are looking at the torsion angle values. While β-turn analysis checks if the chosen torsion angles are fitting the theoretical range for particular β-turn type, Dash cluster analysis is accounting for the similarities in the torsion values across a set of torsion angles (see Section 3.5.1). The details of each analysis method together with the steps taken to obtain data are explained below.

4.6.1 β-turn and hydrogen bond populations

First, the β-turn and hydrogen bond populations in the trajectories were examined.

To analyse **β-turn populations**, specific torsion angles were extracted using the *cpptraj* module of the Amber program, and their values were examined against the β-turn angle ranges (see Section 1.2.1 for β-turn classification description). Here, the script was made to test if the torsion angle values are within the specific range for the particular β-turn. The results of β-turn analysis for AVP simulation repeats are summarised in Table 4.1.

The data show that AVP is preferably adopting conformations with β-turns centred at residues 3,4 and 4,5. The 3,4 β-turn I has population of 26 to 36 %, while 3,4 type II is between 6 - 10 %. 4,5 β-turn II is populated 10 % across all simulation repeats.

Simulation	2,3 type I	2,3 type II	3,4 type I	3,4 type II	4,5 type I	4,5 type II	7,8 type I	7,8 type II
Clinched Open	0.71	0.92	35.45	9.57	10.16	1.35	2.15	5.26
Twisted Saddle	0.87	1.94	26.58	10.13	10.30	0.87	2.26	5.45
Open	1.15	1.46	29.13	6.49	10.37	1.93	1.84	6.40
Saddle	0.63	1.29	35.88	7.53	9.74	0.79	2.33	5.83

Table 4.1: β-turn type populations in the four REMD simulations

The **hydrogen bond populations** across the ring involving the backbone amide bonds were also examined with the help of the *cpptraj* Amber module. A hydrogen bond was defined to exist if the distance between H – O atoms was within 1.6 to 2.4 Å range and the angle between N-H-O atoms is in the 90 - 180° range (see Section 1.2.3). Again, the script was made to check that the conditions imposed were satisfied. The result of the different hydrogen bond populations are given in the Table 4.2.

Hydrogen bond	Clinched Open	Twisted Saddle	Saddle	Open
$Cys^1O - Gln^4H$	0.87	1.54	0.71	1.16
$Tyr^2O - Gln^4H$	2.97	4.32	3.55	3.35
$Tyr^2O - Asn^5H$	53.06	43.14	49.85	41.04
$Tyr^2O - Cys^6H$	39.87	30.94	38.01	30.87
$Phe^3O - Asn^5H$	4.73	6.32	4.08	3.48
$Phe^3O - Cys^6H$	3.18	3.69	2.41	3.28
$Gln^4O - Cys^6H$	4.36	4.81	4.36	5.84
$Cys^6O - Gly^9H$	7.12	7.68	8.22	7.19

Table 4.2: Different hydrogen bond populations from the four REMD simulations.

As can be observed, $Tyr^2O - Phe^5H$ and $Tyr^2O - Cys^6H$ are two the most populated hydrogen bonds in the simulations. Their values vary between 40 to 55 % for $Tyr^2O - Phe^5H$ and 30 to 40 % for $Tyr^2O - Cys^6H$. This suggests that during half of the simulation time AVP conformations were stabilised by the $Tyr^2O - Phe^5H$ hydrogen bond. Other intracyclic hydrogen bonds (residues 1-6) are populated much less and their populations are not as significant as for these two bonds. A hydrogen bond between the ring and tail residues $Cys^6O - Gly^9H$ was also observed with the average population of 7 %.

Summary. The analysis on the β-turns and hydrogen bonds confirmed previous experimental and computational data. There is a conformational preference for AVP to adopt 3,4 and 4,5 β-turns stabilised with $Tyr^2O - Phe^5H$ and

$Tyr^2O - Cys^6H$ hydrogen bonds. A tail was described as adopting 7,8 β-turn in the MD simulation [173] (see Section 4.2). In our simulations, this turn was populated 5 to 10 %, further stabilised with $Cys^6O - Gly^9H$ hydrogen bond in the same population range.

4.6.2 *cis/trans* proline amide bond

Since proline residue is presented in AVP sequence, the *cis/trans* population of the amide bond involving proline nitrogen was also analysed. The rotation around $CA_{cys} - C_{cys} - N_{pro} - CA_{pro}$ bond is defined with ω angle. The *cis* bond was taken as adopting range of +/- 60° from the mean value of 0°. The trans configuration of the ω angle is defined by taking the values at around +/- 180°.

The results are given in Table 4.3, in which it can be seen that *cis* bond conformation is present in two out of four simulation runs. This suggests that maybe the highest temperature in the REMD run (550 K) was not maybe high enough to observe *cis/trans* isomerisation.

In the NMR experiments, the population of the *cis*-proline amide bond was approximately 5 % in one experiment [167], and approximately 9 % in the another NMR experiment [179]. The *cis*-proline isomerisation was not observed in the published MD simulation [173].

amide conformation	Saddle	Twisted Saddle	Open	Clinched Open
cis	1.93	2.29	0.0	0.0
trans	98.07	97.71	100.00	100.00

Table 4.3: The populations of the *cis/trans* amide bonds during the REMD simulations.

4.6.3 Cluster populations

The published MD study reported the AVP conformational ensemble in terms of the adopted ring conformations [173]. They identified four distinct AVP conformational states, named as *Open, Saddle, Clinched Open* and *Twisted Saddle*, characterised with distinct β-turn centres and hydrogen bonds (Section 4.2). Each cluster state was also associated with the distinct $\phi\psi$ torsion angle values of the ring residues.

Following this idea, we also used the same torsion based clustering approach implemented in the Dash software (Section 3.5.1) to assign the AVP conformational ensemble into a number of states.

The identification of the AVP conformational states based on the conformations adopted by the ring part of the structure is commonly used for cyclic peptides, and the published experimental data also follow such a classification.

4.6.3.1 Definition of each cluster state

The torsion time series trajectory from the simulations was obtained extracting the $\phi\psi$ torsion angles adopted by residues Tyr^2 to Cys^6, which were then analysed using the Dash software which groups similar torsions into a number of Dash states (see Section 3.5.1). Given the list of Dash states with associated mean and SD torsion angle values, the Dash states were then compared between themselves, and with the mean values of the states obtained from the MD simulations of Haensele et al. [173]. This part of the analysis was done using *Dashsim* program (see Section 3.5.2). In total, a unique combination of ten torsion angles defines each cluster state.

As a result, four distinct cluster states were obtained with the REMD simulations (Figure 4.6). Each cluster state was defined based on the mean and standard deviation of the angles of the ring residues (see Appendix A). The representative cluster states cover approximately 70 % of the REMD simulations. The remaining conformations were defined as transient states because they could not be assigned to any of the cluster states, and showed no overlap between the REMD simulation repeats.

The cluster states reported from the REMD simulations are the same states as were reported in the MD simulation. The comparison between the MD and REMD ring torsion mean and SD values is given in the next Section in Figure 4.12.

4.6.3.2 Cluster state population time evolution

The simulations were then analysed in such a way that each simulation trajectory was divided into 100 equal parts, and for the each part it was calculated the population of each cluster states. The idea is shown in Figure 4.7, where one of the simulation trajectories is shown to be divided into several parts just to describe the idea of the analysis. In each part, the population of each of the cluster states is calculated, and colour coded to show the time evolution of the particular cluster

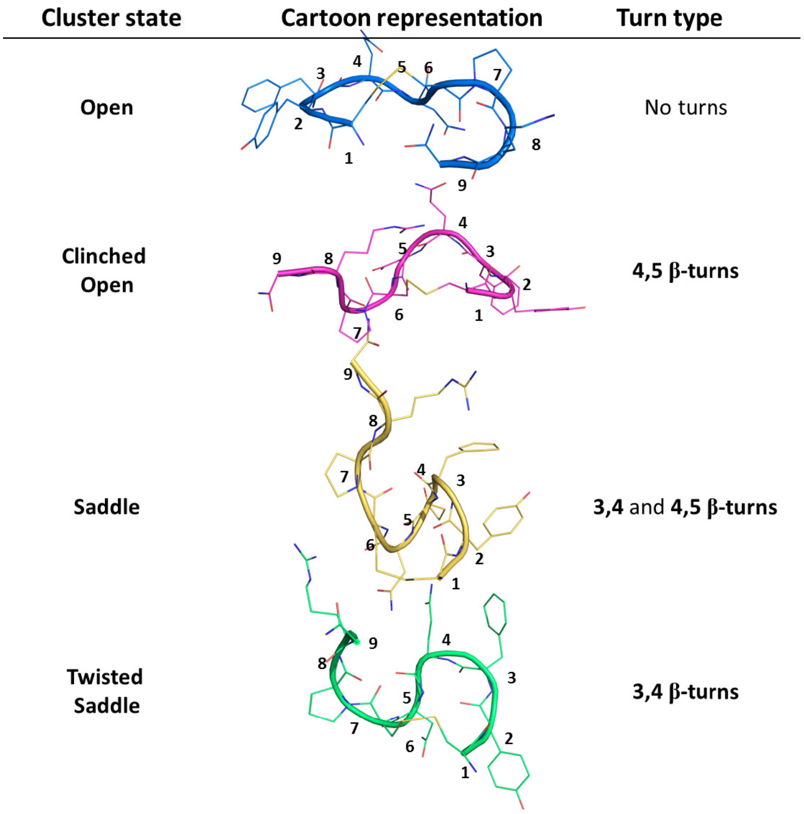


Figure 4.6: The cluster states adopted by AVP with associated turn types.

state. A Dash state was identified as to belong to the particular cluster state if the calculated similarity score was higher than 0.65. The threshold value was obtained from RMSD analysis on the backbone atoms of the cyclic part of the structure where the minimal RMSD similarity value between two cluster states corresponds to 65 % similarity score (Table 4.4).

	Saddle	**Twisted Saddle**	**Open**	**Clinched Open**
Saddle	**0.00**	0.30	0.47	0.35
Twisted Saddle	0.30	**0.00**	0.31	0.32
Open	0.47	0.31	**0.00**	0.33
Clinched Open	0.35	0.32	0.33	**0.00**

Table 4.4: RMSD scores between different AVP cluster states

Following this approach, all the simulation repeats were analysed in terms of the cluster state population as a function of simulation time.

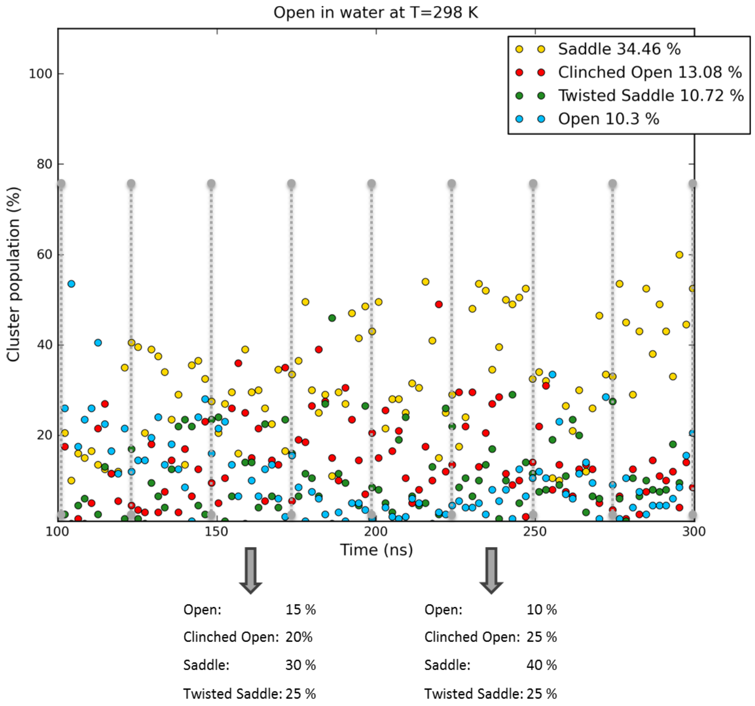


Figure 4.7: The analysis of the cluster state evolution is shown for the example of the REMD *Open* trajectory, which is divided into equal parts, and in each part the population of the individual cluster state was calculated. For example, in one of the chunks it was calculated with the Dash that cluster state populations are following: Open 15 %, Clinched Open 20 %, Saddle 30 % and Twisted Saddle 25 %. These are the approximative populations just to show the idea used to get the population plot. These data are then used in the population plot, Figure 4.8

4.6.3.3 Results

The final populations of all REMD simulation repeats are shown in Figure 4.8. Irrespective of the starting conformation, the *Saddle* state seems to be the most populated. All the cluster states, but *Saddle*, vary their populations between a lower value and 30 %, while *Saddle* cluster keeps the population 30 to 60 % through the all simulation repeats.

Another way to show the convergence of the cluster population ratios was to plot the **cumulative averaged population** of each cluster state to see the population ratios settle down (Figure 4.9). The cumulative averaged cluster population values match very well with the overall Dash cluster populations given in figure legend, so these population ratios were used in further analysis.

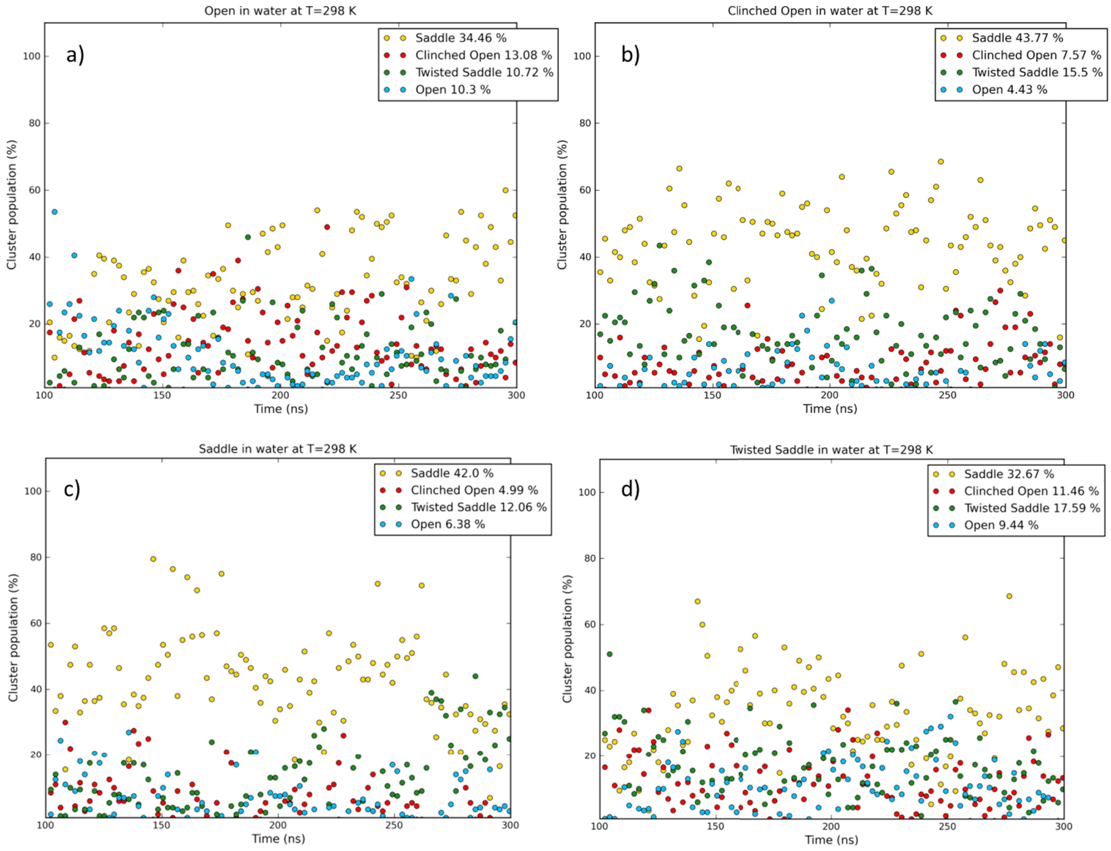


Figure 4.8: The population of the individual cluster states from four REMD simulations, each run with different starting conformation: *a) Open*, *b) Clinched Open*, *c) Saddle*, *d) Twisted Saddle*.

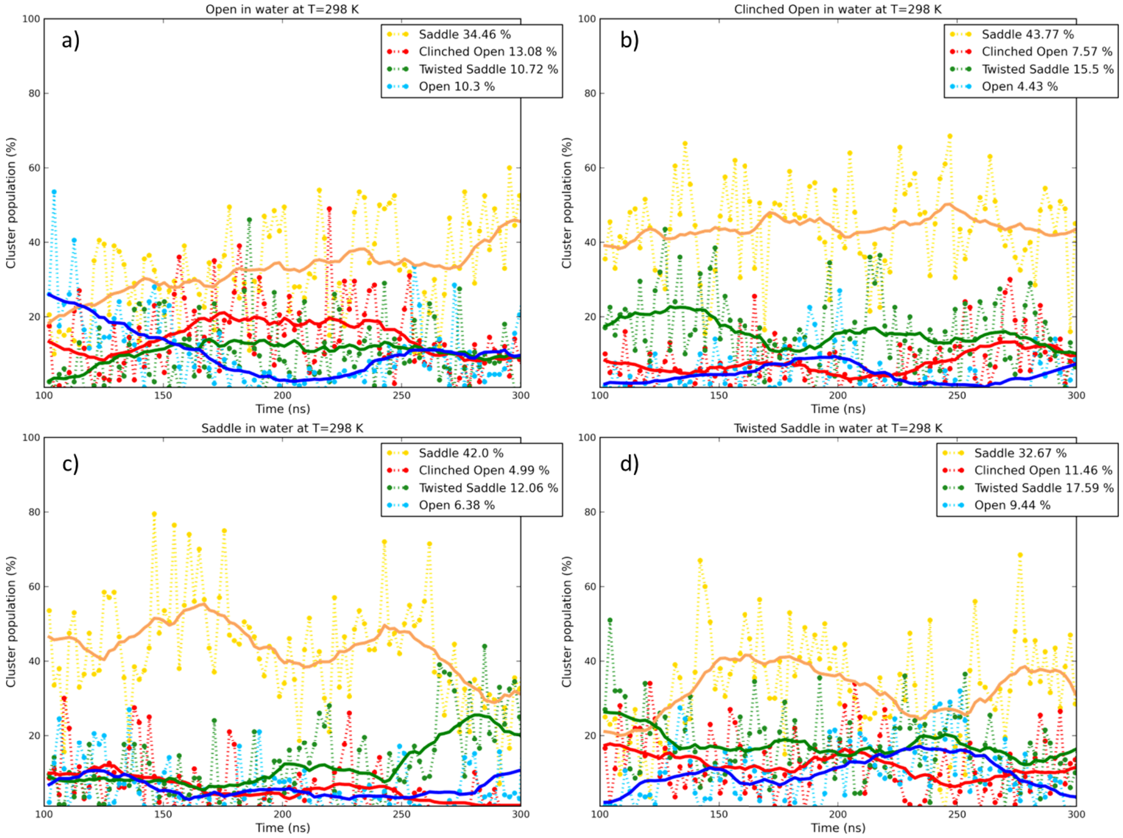


Figure 4.9: The population of the individual cluster states from four REMD simulations, each run with different starting conformation: *a) Open*, *b) Clinched Open*, *c) Saddle*, *d) Twisted Saddle*. The bold lines show the cumulative averaged population of the particular cluster state.

4.6.4 The population of *cis* amide in AVP cluster states

Table 4.3 reports that in REMD simulations the *cis*-proline amide bond appeared with a population of approximately 2 %. Moreover, now that the AVP cluster ensemble is defined, the *cis* isomer was examined to see if it is localised in any of the AVP cluster states (Table 4.5).

Figure 4.10, together with Table 4.5 shows that *cis* isomer is mostly populated by the *Saddle* conformation, while less presented in other conformations. The *cis*-proline isomer was also reported in NMR studies [167, 168], but they also reported folded (*Saddle*) ring conformation only, so it needs to be taken with caution that only a folded ring conformation adopts a *cis* isomer in C terminal tail. Our data suggest that cis isomer is not necessary system specific, but the highest temperature in REMD was not maybe high enough to equilibrate across the 20.6 kcal/mol energy barrier for AVP Pro *cis/trans* isomerisation [179].

A detailed population of *cis* amide bond in each cluster state is given in Table 4.5.

Cluster state	REMD Saddle	REMD Twisted Saddle
Saddle	41.93	40.55
Twisted Saddle	16.41	12.82
Clinched Open	5.38	4.62
Open	3.52	1.48

Table 4.5: The population of *cis*-proline amide bond in each AVP cluster state in two REMD simulations. The data were obtained from the analysis of the T=298 K trajectory.

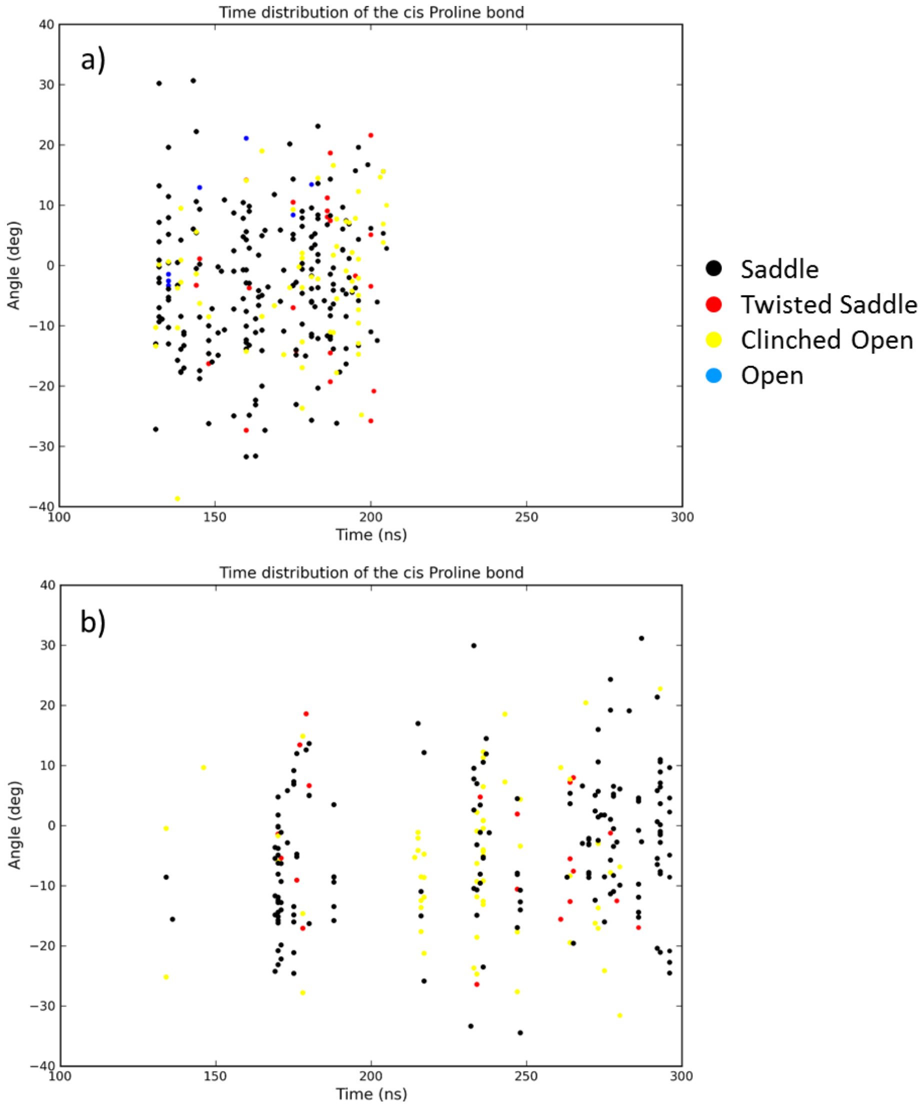


Figure 4.10: The population of *cis*-proline amide bond in the AVP a) REMD Twisted Saddle, and b) REMD Saddle simulations.

4.6.5 Summary

The initial approach to obtain a converged AVP conformational ensemble was achieved using the enhanced sampling REMD method (Figure 4.8), although the *cis*-proline amide populations may not be converged (Table 4.3). Four simulation repeats showed that AVP is interconverting between four cluster states (*Open, Saddle, Clinched Open* and *Twisted Saddle*), with *Saddle* being the preferred structure. The assigned cluster states were populated similarly in all simulation runs (Figure 4.8).

The AVP conformational ensemble members were named according to the previously published MD study, which used the same clustering software to assign conformational states as we did. Dash was chosen as a method of choice to check for the conformational diversity because it assigns together the frames adopting similar torsion values .

REMD conformational ensemble. The population plot of the AVP cluster states during the simulation time (Figure 4.8) shows that *Saddle* cluster state averages around 40 % during the simulation time in any one particular simulation block. The second most populated structure is *Twisted Saddle*, between 10 and 20 %, followed by *Clinched Open* with the same approximate population of 10 to 20 %. In total, the three most populated structures account for 60 to 80 % of the AVP conformations. The *Open* state is the lowest populated state with population of approximately 10 %. The remaining 30 % of the total populations were considered as transient states because they had very low populations across simulation repeats.

Comparison between cluster state populations and β-turn/hydrogen bond populations. The overall β-turn population of 45 to 55 % (Table 4.1) agrees well with the total population of *Saddle, Twisted Saddle* and *Clinched Open* cluster states, 60 to 80 % (Figure 4.8). Hydrogen bonds which are thought to stabilise ring conformations of these three cluster states are also highly populated between 40 and 55 % (Table 4.2). Therefore, the overall population of adopted β-turns and hydrogen bonds in the simulation trajectories are showing similar populations to cluster state populations. An *Open* state is only possible to assign from the torsion angle analysis since it is not stabilised by any hydrogen bond and does not adopt any turn types, so we cannot say from the turn and hydrogen bond analysis that anything which does not contain a turn or hydrogen bond is the open cluster state. Because of that, its population cannot be double checked, so it is taken that it is populated approximately 10 %.

Comparison between MD and REMD data. The published MD simulations [173] reported the same cluster diversity as we did, but the population of the states differ. The *Open* starting cluster conformation was populated 13 %, followed by 40 % of the *Saddle* conformation and *Clinched Open* populated 7 %. The last observed state in the MD simulation trajectory was *Twisted Saddle* with 34 %. Other conformations were taken as variants, and they were populated 6 % during the 11 μs MD simulation time. The results obtained in this MD study are shown in Figure 4.4. As can be seen from the figure, the dynamics of the interchanging states was very slow, so the populations of the cluster states cannot be considered as converged. In this study, the tail was characterised as *extended*, or *folded* if the 7,8 β-turn was populated in the C terminal tail. The approximative population ratio reported for the *extended* vs. *folded* tail was 80 % : 20 %. In our simulations, a 7,8 β-turn was less populated, between 5 and 10 %, further stabilised with a hydrogen bond between the ring residue Cys^6 and the tail residue Gly^9 in the same population range.

Although the AVP cluster states from our simulations and MD simulation seem to overlap, it can be claimed that REMD produces a converged conformational ensemble compared to MD which observed very slow interconversion between the cluster states. The choice of the temperature range (Figure 4.5) is sufficient to overcome energy barriers between local minima, but not high enough for consistent *cis/trans* amide bond isomerisation (Table 4.3). Overall, REMD provided a good description of the AVP conformational ensemble.

4.7 AVP chemical shifts

Next, it was decided to validate the AVP conformational ensemble against NMR chemical shift data [180]. In order to do so, the chemical shifts were calculated for the AVP structures obtained from the REMD simulations.

The proton chemical shifts were calculated using Gaussian09 software [116] with B3LYP/6-31G(d) level of DFT theory (see Section 3.2). This choice of functional and basis set has been shown to be appropriate to calculate chemical shifts of biomolecules [181].

4.7.0.1 The choice of the representative structures

The AVP was recognised to adopt four different ring conformations *Open, Clinched Open, Saddle* and *Twisted Saddle.* For each cluster state, several structures were extracted to fulfil the following conditions:

- the structures are scattered approximately in equal parts along the trajectory
- the ring torsions of the chosen structures are within 1 SD of the torsion angle distribution for that cluster state

The first conditions was chosen to make sure that the structures are taken from the different parts of the trajectory (Figure 4.11), while the second condition meant that chosen structures are truly representative of the cluster state (Figure 4.12). *Saddle, Twisted Saddle* and *Open* are presented with 8 structures, while *Clinched Open* with 6. All the representative structures were taken from the AVP *Clinched Open* simulation.

4.7.0.2 Chemical shift calculation protocol

First, using *babel* the structures were converted into *mol2* format, and then into Gaussian Z-matrix (*gzmat*) format. Here, it was specified the level of theory desired with appropriate basis set, structure specifications including charge and spin state, inclusion of PCM water model, and some output file details. Chemical shift calculation was performed in two steps:

1. Each structure was optimised at the specified level of theory

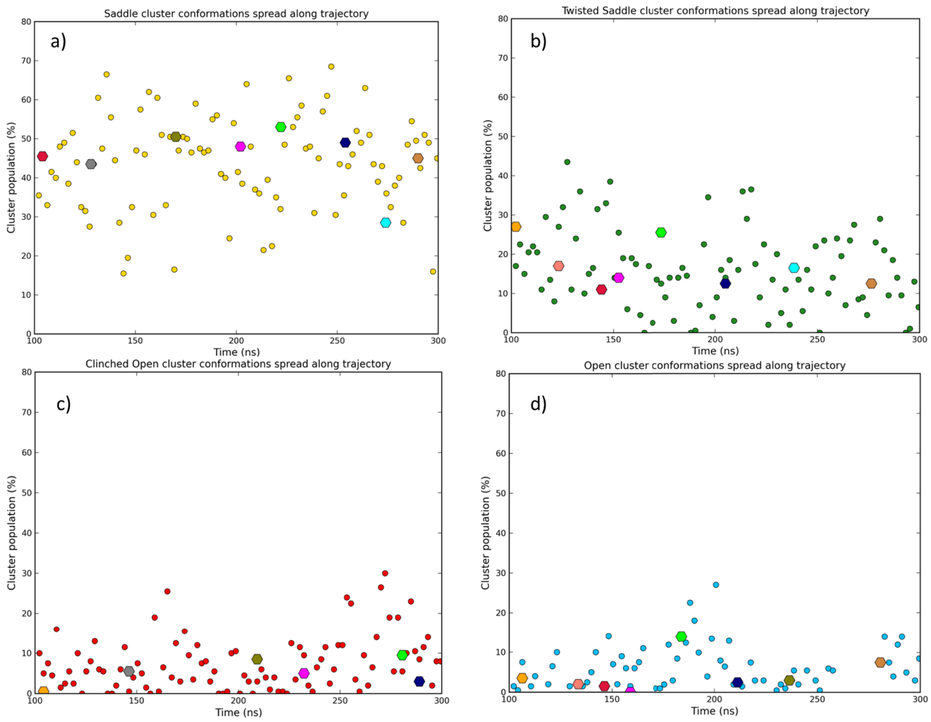


Figure 4.11: The locations of the pulled out representative structures for each cluster state a) Saddle, b) Twisted Saddle c) Clinched Open, d) Open during the simulation time. The emphasized colour dots depicts another structure for given cluster state.

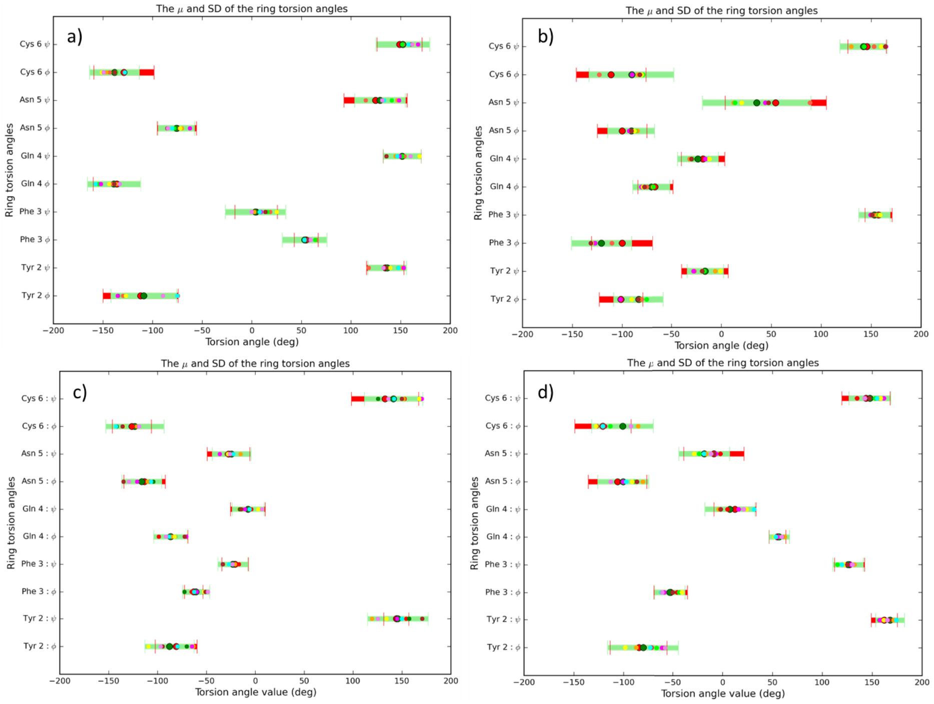


Figure 4.12: The distribution of observed torsion angles for each conformation in each conformational state for a) Saddle, b) Twisted Saddle, c) Clinched Open, d) Open. The red bars are from MD simulations [182], and REST torsion angle distributions are in green. The spots show the dihedral angles for the structures we selected.

2. Shielding constants for each atom in the structure were calculated, and later converted into chemical shifts. Chemical shifts for each optimised cluster conformation were calculated using the regression equation [117]

$$\delta(^1H) = -0.9912\sigma_H + 32.05 \tag{4.1}$$

where δ is the chemical shift and σ calculated isotropic atomic magnetic shielding constant [117, 180].

The values of the calculated chemical shifts were then compared with the experimental values obtained at pH 6.2 and temperature 298 K [180].

4.7.0.3 Comparison of the calculated and experimental chemical shift data

Since each cluster state is represented by a few structures, and for each of them proton chemical shifts were calculated, first the calculated chemical shifts were analysed. In total, there are experimentally reported values for 35 proton shifts [180]. The calculated proton shifts were extracted to match the number of experimental proton shifts. The statistical analysis of the calculated and experimental chemical shifts consisted of a few parts:

- the variance within each calculated chemical shifts type
- the goodness of fit between the calculated and experimental chemical shifts
- the intra-cluster R^2 distribution of the calculated vs. experimental chemical shifts
- the inter-cluster R^2 distribution of the calculated vs. experimental chemical shifts

The variance within each chemical shift type between the representative structures belonging to each identified AVP cluster was calculated to check which chemical shift types are adopting larger variance compared to others within the same cluster and between the AVP cluster types. Figure 4.13 shows that the chemical shift types Tyr^2 HE*, Phe^3 HZ, Asn^5 HB*, Pro^7 HD2, HD3, HG*, Arg^8 HG*, HD* and Gly^9 HA are adopting the tightest range in all AVP cluster states with the variance between the representative chemical shift types less than 0.05

ppm. The chemical shift types with the widest variance belong to Cys^1 and Arg^8 residues (light colours in Figure 4.13). The star next to the chemical shift type denotes that the particular chemical shift type was identified only as the averaged signal over two protons attached to the same heavy atom.

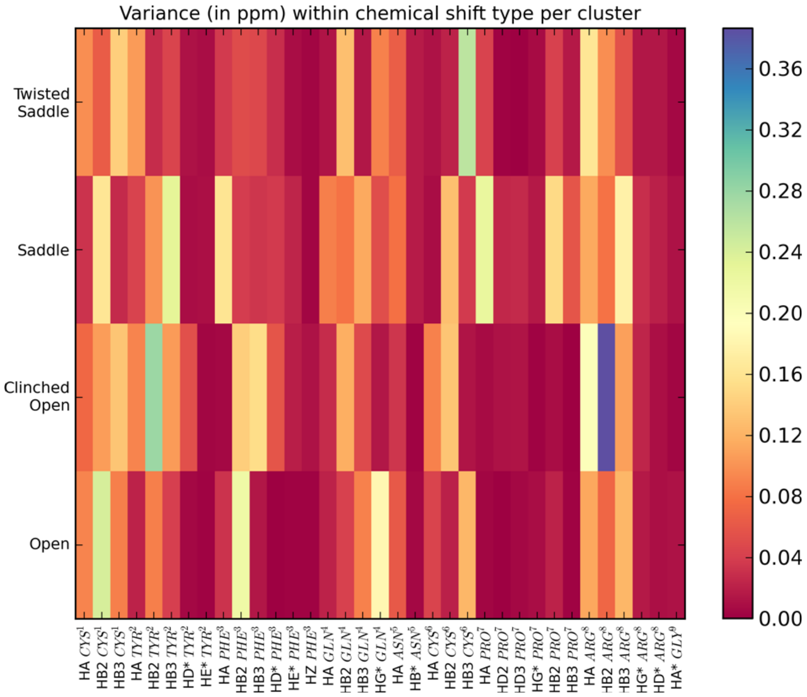


Figure 4.13: The chemical shift type variance within each AVP cluster.

There is a linear fit between the calculated and experimental chemical shifts for all AVP cluster states which are adopting similar distributions of the chemical shift values against experimental data (Figure 4.14). Each theoretically calculated proton chemical shifts type is taken as the mean with associated SD on the error bars.

The intra-cluster variance analysis was another type of the analysis performed to obtain an indication of the difference between the extracted individual representative structures within particular AVP cluster state in terms of R^2 values (Figure 4.15).

The R^2 values show that the best agreement with experimental data have *Clinched Open* and *Saddle* cluster state structure chemical shifts. They were all distributed in the range 0.95 to 0.97. The widest R^2 distribution adopt *Twisted Saddle* representative structures, from 0.925 to 0.96. However, the overall R^2 range of the individual structures is very similar across all cluster states.

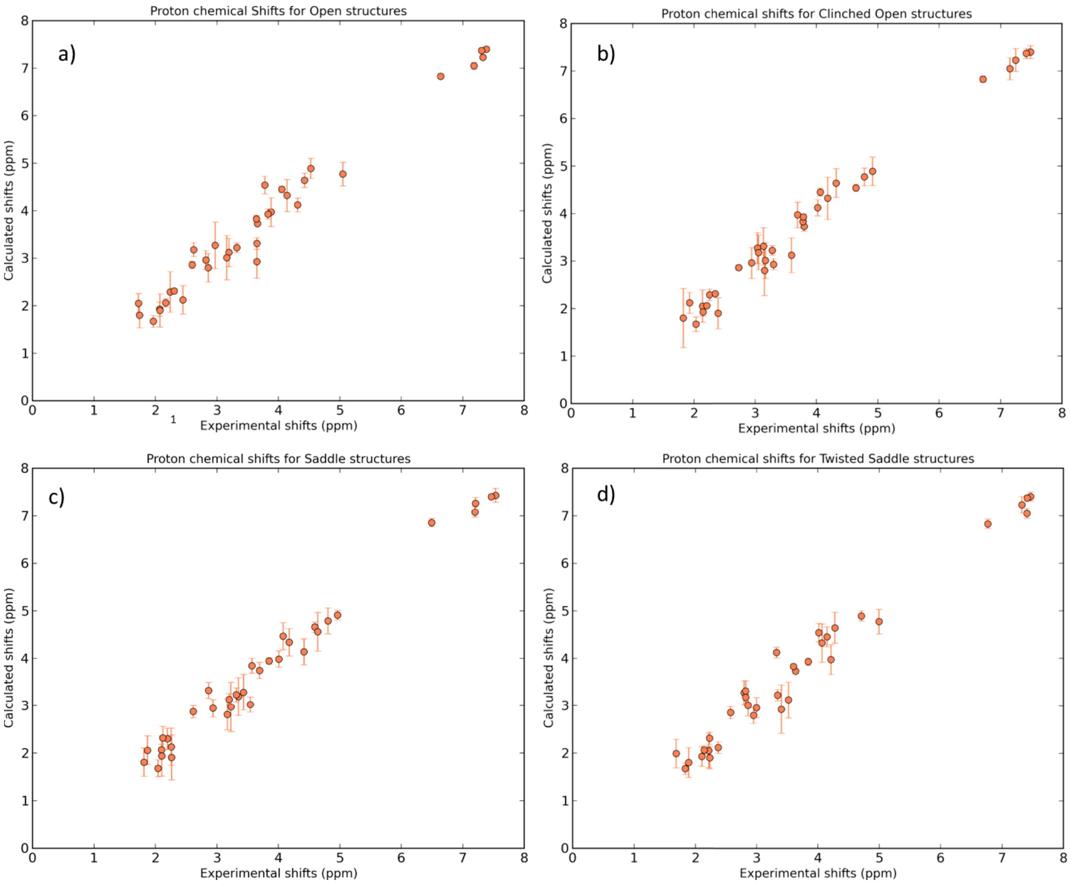


Figure 4.14: The comparison between the experimental chemical shifts and computational calculated chemical shifts given as the average with the standard deviation as error bars for the a) Open, b) Clinched Open, c) Saddle and d) Twisted Saddle representative structures.

4.7.1 Bootstrapping of the individual cluster states

Since, for computational reasons, only a few structures were chosen to represent a particular cluster state, the chemical shift data were then bootstrapped with 95 % CI to check how each cluster state would behave irrespective of the chosen structures. The bootstrapped R^2 distribution is given in Figure 4.16. Here it is shown that the best performing structures are associated with the *Saddle, Clinched Open* and *Open* clusters, while *Twisted Saddle* has lower R^2 value; the mean is centred at 0.950. *Clinched Open* and *Saddle* take almost identical mean values, 0.959 and 0.958, respectively.

Besides bootstrapping R^2 values of the chemical shifts, other statistical measures of similarity, MUE and MSE were also bootstrapped. Their values are given in Table 4.6.

The bootstrapped values of statistical metrics used to compare computational and experimental chemical shifts show that the best agreement is for *Clinched*

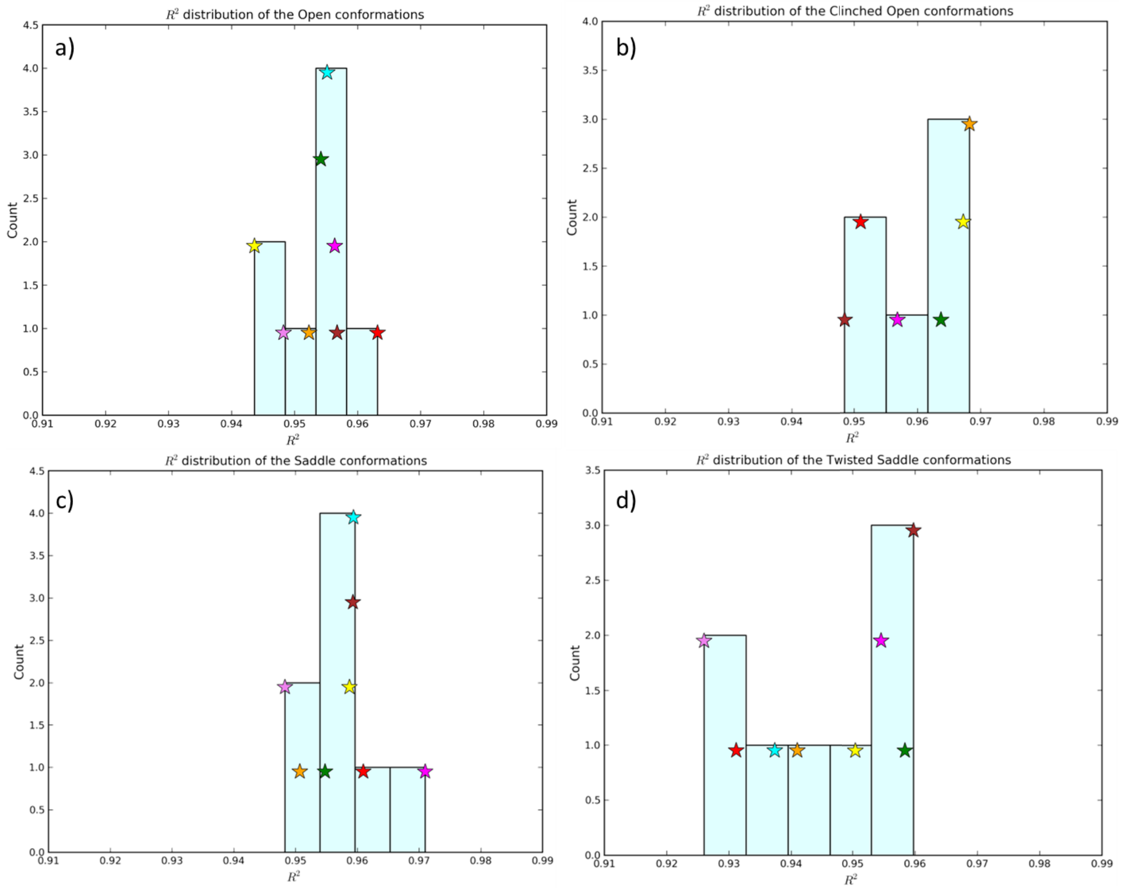


Figure 4.15: The distribution of the R^2 values calculated between experimentally measured and theoretically obtained chemical shifts for each structure belonging to particular AVP cluster state: a) Open, b) Clinched Open, c) Saddle, d) Twisted Saddle. The asterix depicts the R^2 values for each representative structure, and were vertically offset to show their spread within a bar.

	MUE	MSE	R^2
Open	0.186 < 0.239 < 0.303	-0.166 < -0.054 < 0.257	0.933 < 0.954 < 0.975
Clinched Open	0.131 < 0.189 < 0.249	-0.105 < -0.018 < 0.068	0.938 < 0.959 < 0.980
Saddle	0.164 < 0.201 < 0.289	-0.074 < 0.023 < 0.012	0.938 < 0.958 < 0.977
Twisted Saddle	0.149 < 0.221 < 0.272	-0.179 < -0.079 < 0.021	0.927 < 0.950 < 0.974

Table 4.6: The bootstrapped values of three statistical measures of similarity, MUE, MSE and R^2 (see Section 3.5.3) for four AVP cluster states.

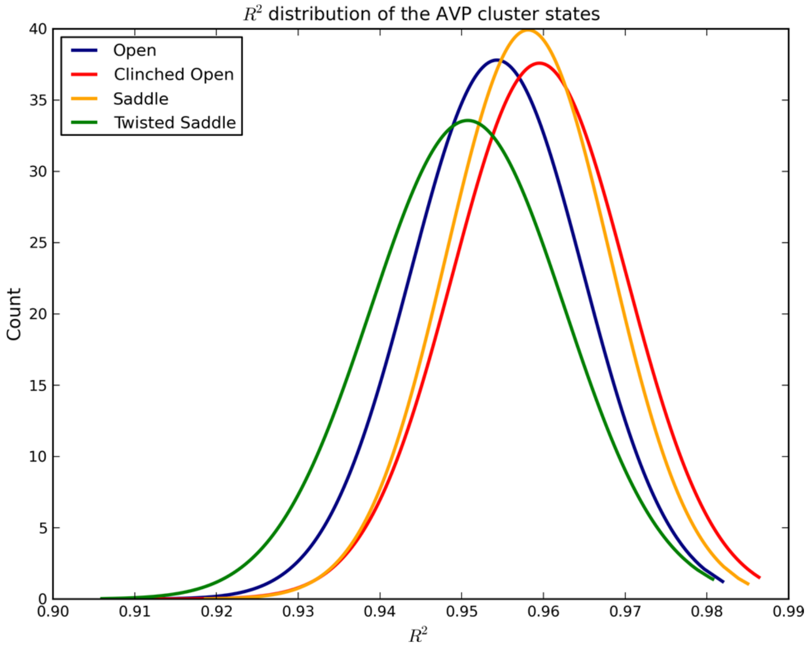


Figure 4.16: Distribution of the bootstrapped R^2 chemical shift values for each AVP cluster state.

Open and *Saddle* cluster states. This was confirmed by MUE and R^2 values.

4.7.1.1 Analysis of the individual chemical shift types

To check which chemical shift types are giving more weight to the final R^2 distribution, the chemical shift types with the variance smaller than 0.005 ppm, 0.01 ppm, 0.02 ppm, 0.04 ppm and 0.06 ppm per shift type were extracted from the sample data on which the statistical analysis was then performed.

By extracting chemical shift types with the lowest variance, it was possible to identify individual chemical shift types with the highest weight to the R^2 distribution when compared with the experimental data. Figure 4.17 (a) shows that lowest weight to the R^2 distribution for the *Open* cluster state comes from the chemical shift types depicted in red and yellow, belonging to the residues Tyr^2, Phe^3, Asn^5 and Pro^7. All the Pro^7 chemical shift types but HB2 are showing very tight range of the values with the variance smaller than 0.01 ppm.

Figure 4.17 (b) shows different R^2 distribution for different chemical shift lists for *Clinched Open* cluster type. Chemical shift types with the variance lower than 0.06 ppm give smaller weight than those higher than 0.06 ppm (orange line).

In case of the *Saddle* cluster (Figure 4.17 (c)), there is only one chemical shift type Phe^3 HZ which has variance smaller than 0.005 ppm compared to all other cluster types. Here chemical shift types with the variance smaller than 0.04 ppm (shown in green) together with three chemical shift types Tyr^2 HA, Gln^4 HG* and Pro^7 HB3 (shown in orange) give very similar weight to the overall R^2 distribution.

Finally, the R^2 distribution for the *Twisted Saddle* cluster type (Figure 4.17 (d)) shows similar pattern as the *Saddle* R^2 distribution where the highest weight to the overall distribution comes from the chemical shift types with the variance higher than 0.04 ppm.

To summarise, chemical shift types with the variance already higher than 0.01 ppm give higher weight to the overall R^2 distribution for *Open* cluster, while for *Clinched Open* cluster state, these are the chemical shift with variance > 0.06 ppm. For *Saddle* and *Twisted Saddle* cluster types, the highest weight comes from the chemical shift types with the variance higher than 0.04 ppm.

Having considered the performance of each cluster in terms of reproducing the experimental chemical shifts, and analysing the weights of the individual chemical shift types, the performance of the simulation ensemble as a whole will be assessed.

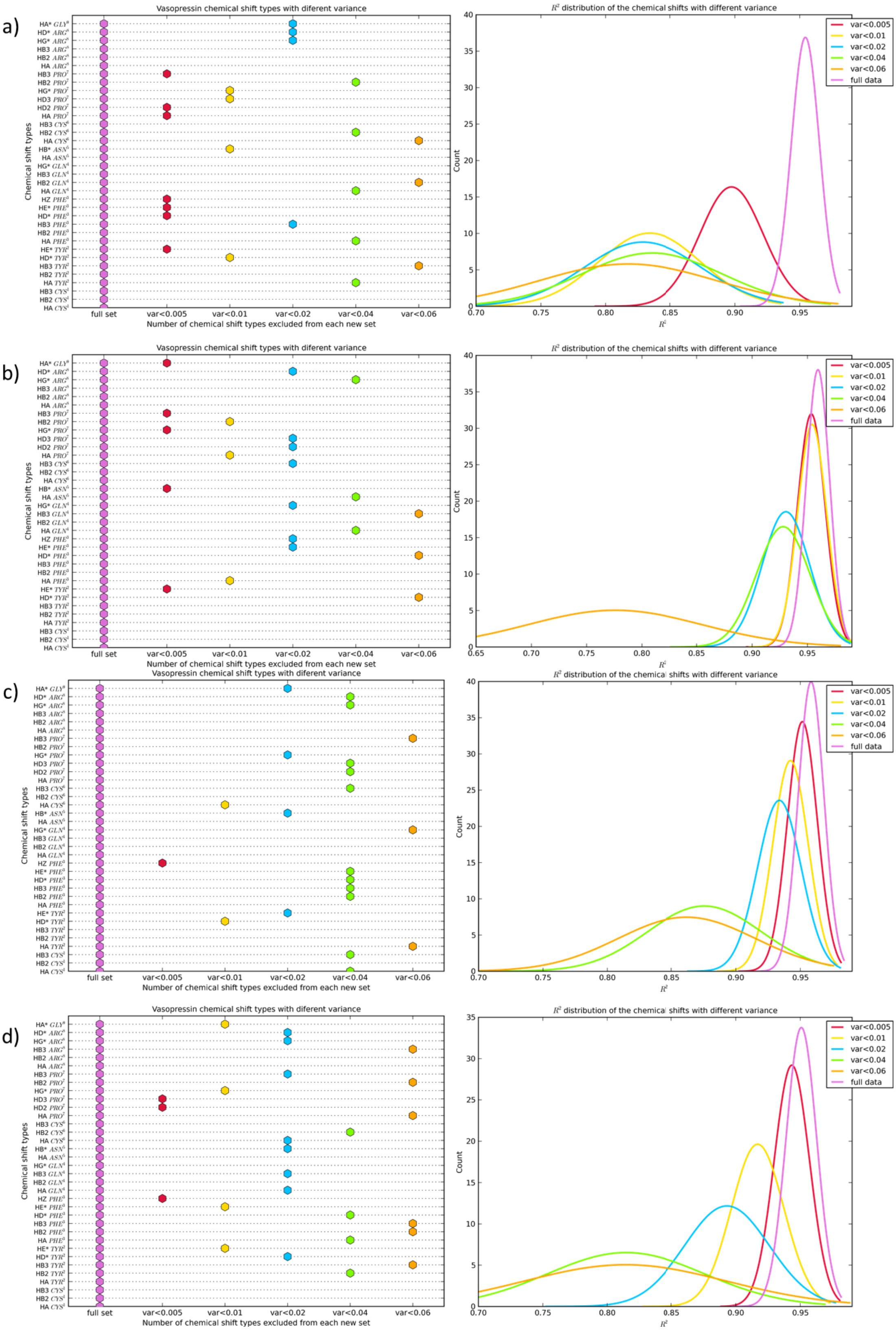


Figure 4.17: The R^2 distributions of the theoretically calculated vs. experimental chemical shifts with different chemical shift data sets depending on the value of their variance (on the left plots). Distributions are plotted separately for a) Open, b) Clinched Open, c) Saddle and d) Twisted Saddle AVP cluster state.

4.7.2 Ensemble model

To validate simulation data against experimental, the idea of the ensemble model was introduced in the recent paper by Haensele at al. [180]. The model is built in such a way that the shift values from structure representative of each cluster state are weighted by the population of the cluster state. The weighted shift values are summed over all cluster states for a particular shift type.

The ensemble model for the AVP peptide was built following the procedure given in the flowchart in Figure 4.18. The same procedure was taken for all peptides studied in this work. The steps taken to obtain the ensemble model are as follows:

1. Proton chemical shifts were calculated for representative structures from each of the AVP cluster states (*Open, Clinched Open, Saddle, Twisted Saddle*). There are 8 representative structures for each of the *Open, Saddle, Twisted Saddle* cluster, and 6 for *Clinched Open* cluster.

2. Shift values belonging to each representative structure were extracted one at a time at random. An example is given on the flowchart where in purple is emphasised the structure which was randomly selected from each cluster state.

3. The selected shift values from each cluster state were then multiplied with associated normalised cluster state population. Here is given the example of the cluster populations from the REMD *Open* simulation run where the normalised population of the cluster states were Open 14 % - Clinched Open 16 % - Saddle 53 % - Twisted Saddle 17 %. The ensemble model was tested on the populations from all four REMD runs.

4. This procedure of randomly selecting shifts belonging to particular structure was repeated 10 000 times. Following this procedure the R^2 metrics with 95 % confidence intervals for the error was built.

The results for the ensemble model obtained following the described protocol are given in Figure 4.19.

The bootstrapped R^2 distribution for the AVP ensemble models shows that the best agreement with experimental data is derived from the ensemble created using REMD dervived from the *Twisted Saddle* starting structure followed by REMD

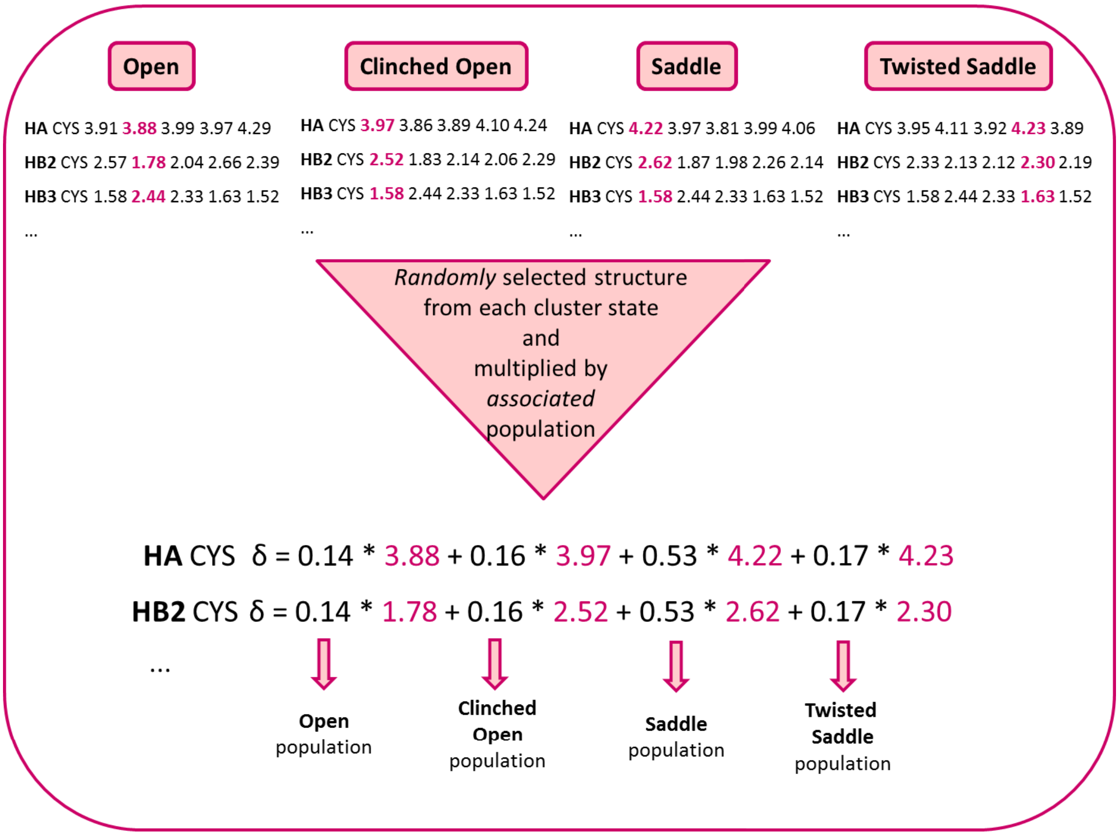

Open
Clinched Open
Saddle
Twisted Saddle
HA CYS 3.91 3.88 3.99 3.97 4.29
HB2 CYS 2.57 1.78 2.04 2.66 2.39
HB3 CYS 1.58 2.44 2.33 1.63 1.52
...
HA CYS 3.97 3.86 3.89 4.10 4.24
HB2 CYS 2.52 1.83 2.14 2.06 2.29
HB3 CYS 1.58 2.44 2.33 1.63 1.52
...
HA CYS 4.22 3.97 3.81 3.99 4.06
HB2 CYS 2.62 1.87 1.98 2.26 2.14
HB3 CYS 1.58 2.44 2.33 1.63 1.52
...
HA CYS 3.95 4.11 3.92 4.23 3.89
HB2 CYS 2.33 2.13 2.12 2.30 2.19
HB3 CYS 1.58 2.44 2.33 1.63 1.52
...
Randomly selected structure from each cluster state and multiplied by associated population
HA CYS δ = 0.14 * 3.88 + 0.16 * 3.97 + 0.53 * 4.22 + 0.17 * 4.23
HB2 CYS δ = 0.14 * 1.78 + 0.16 * 2.52 + 0.53 * 2.62 + 0.17 * 2.30
...
Open population
Clinched Open population
Saddle population
Twisted Saddle population

Figure 4.18: The flowchart describes the approach taken to calculate ensemble chemical shifts. The step by step explanation is given in the main text (Section 4.7.2).

ensemble from the *Open* structure. The approximate populations of the individual states are listed below:

- **REMD Twisted Saddle**: *Saddle* 33 % - *Clinched Open* 12 % - *Twisted Saddle* 18 % - *Open* 10 %
- **REMD Open**: *Saddle* 35 % - *Clinched Open* 11 % - *Twisted Saddle* 11 % - *Open* 10 %
- **REMD Clinched Open**: *Saddle* 44 % - *Clinched Open* 8 % - *Twisted Saddle* 16 % - *Open* 5 %
- **REMD Saddle**: *Saddle* 42 % - *Clinched Open* 5 % - *Twisted Saddle* 12 % - *Open* 7 %

The ratio of the populations in the best performing ensemble model (REMD *Twisted Saddle*) shows that the *Saddle* conformation is the most populated conformation in the ensemble, but among the four ensembles it contains the lowest proportion of *Saddle*, while *Twisted Saddle*, *Clinched Open* and *Open* cluster states are the highest populated in this ensemble compared to the other REMD ensembles. In terms of ratios of the individual cluster state population in this ensemble, *Saddle* and *Twisted Saddle* are highest populated, while *Clinched Open* and *Open* are lower, but these two conformers adopt almost equal populations (12 % and 10 %, respectively) (Figure 4.8). *Saddle*, as one of the unique cluster states, has the highest individual R^2 value of all the individual states (Figure 4.16), and it is also the highest populated state in all the simulations, so the contribution of this state to the model is the biggest in terms of weight. Besides *Saddle*, *Clinched Open* state has also interesting behaviour. This state, just like *Saddle*, has the highest R^2 value, but on the other side, it is lower populated in the simulations (5 % to 12 %), so the contribution of this state to the model will be lower than of the *Saddle* conformer.

Bootstrapping of individual cluster state vs. ensemble model. If we compare individual cluster states bootstrapped R^2 (Figure 4.16) with the ensemble bootstrapped R^2 distribution (Figure 4.19), then obviously ensembles match better with experimental data than any separate cluster state.

The best performing ensemble simulation REMD *Twisted Saddle* has R^2 averaged at 0.981 while the best performing individual cluster state *Clinched Open* has the R^2 peak positioned at 0.960. The worst agreement with experimental data from all ensembles is for the REMD *Saddle* and REMD *Clinched Open* ensembles ($R^2 \approx 0.975$). If we compare these values with the best individual cluster state *Clinched Open* which has a peak centred at $R^2 \approx 0.959$ (Table 4.6), than we see that even the worst performing ensemble model has better agreement with experimental data than any individual cluster state. This additionally supports the idea that intrinsically disordered peptides exist in an ensemble of conformations rather than as one structure.

Moreover, if we compare the obtained data with the experimental evidence given in the Section 4.2, then it can be observed that the ensemble model reflects the AVP conformational diversity. The best performing ensemble models have high percentages of the *Open* and *Saddle* cluster state which resemble two AVP crystal structure.

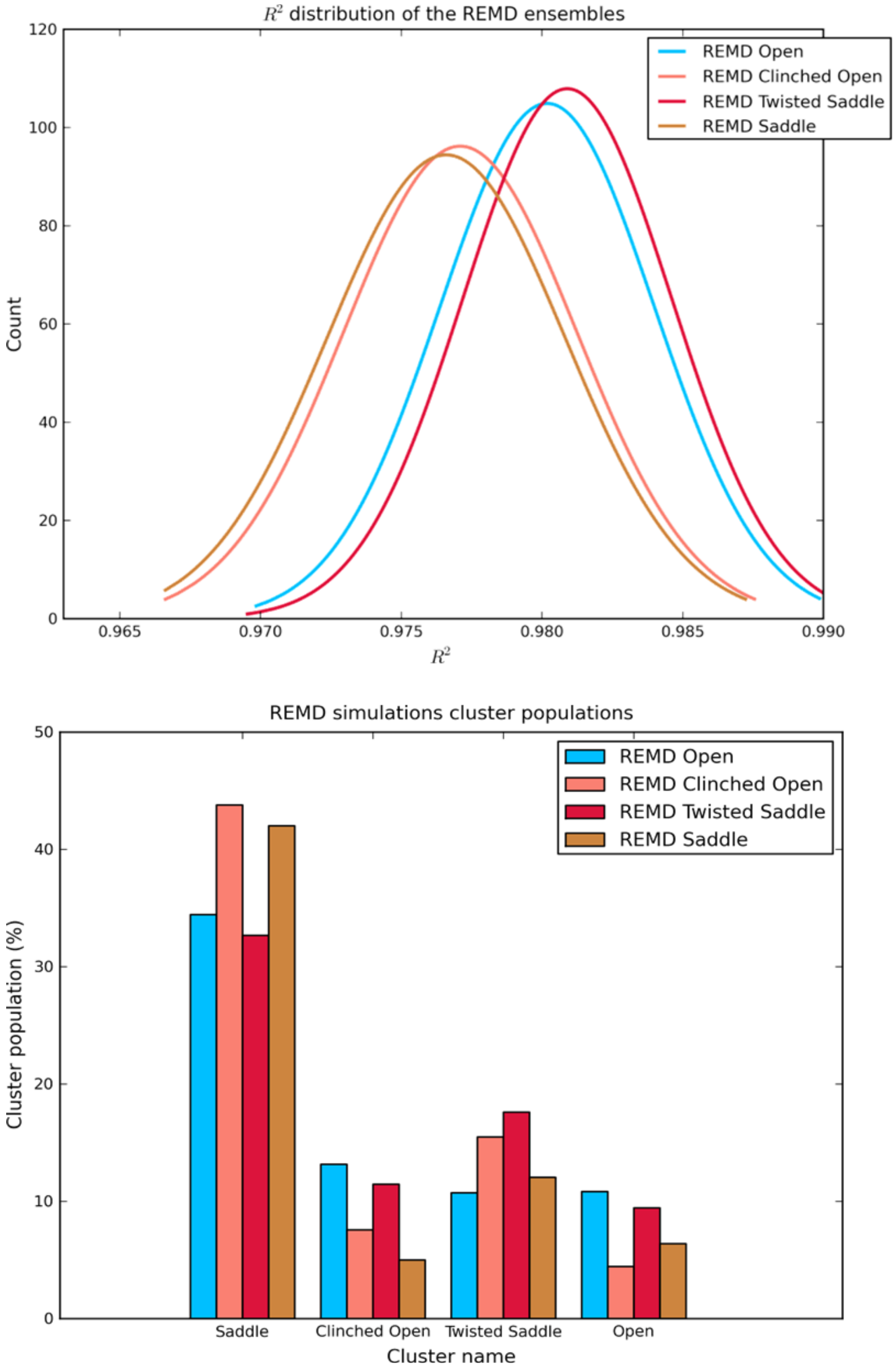


Figure 4.19: The upper part of the picture shows the bootstrapped R^2 distribution of the ensemble model, while the plot below shows the populations of the each cluster state in each of the four simulation repeats. Ensembles derived from each simulation starting structure are colour coded (REMD *Open*, REMD *Clinched Open*, REMD *Saddle* and REMD *Twisted Saddle*).

4.7.2.1 Optimal cluster population ratios

We also wanted to examine the population ratios which yield the best agreement of the calculated with the experimentally measured chemical shifts. The ensemble model was built in the same way as with the simulation cluster populations, but here the cluster populations were generated randomly and only those population ratios which gave the correlation coefficient $R^2 > 0.99$ were kept. The results are given in Figure 4.20.

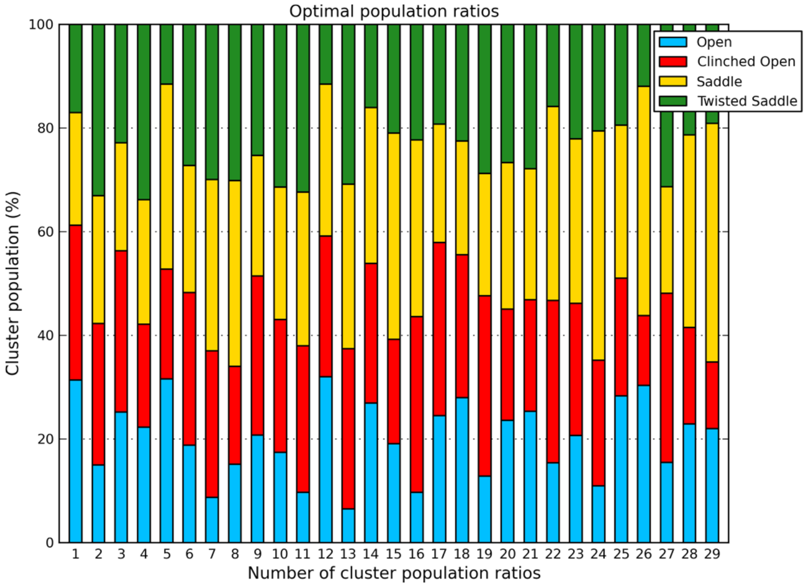


Figure 4.20: The population ratios of the AVP cluster states which yielded correlation coefficient $R^2 > 0.99$ when comparing experimental chemical shift values with the ensemble model.

It shows that the sum of chemical shifts weighted by the different population ratios between the AVP cluster states gives very good agreement with the experimental values. However, there is no preferred population ratio, and different population values of the AVP cluster states give the same result, but the overall population ratio matches the simulation population data with the most populated cluster state *Saddle*, followed by *Clinched Open*, *Open* and *Twisted Saddle*. The results also support initial idea that the ensemble model is better approximation of the AVP conformational diversity than one global structure.

4.8 Conclusions

AVP is a small cyclic peptide whose conformations have been probed with experimental and computational techniques reviewed in Sections 4.2 and 4.3. The REMD simulation data contributed to already published data in such a way that it gave a complete converged picture of the AVP conformational ensemble (Figure 4.8).

These results were then validated by comparison with experimental chemical shift data. Chemical shifts are commonly measured NMR observables. The structures from the reported AVP cluster members were validated against experimental chemical shift values. This analysis revealed that the closest chemical shift values to the experimental data are for the individual cluster states *Saddle* and *Clinched Open*.

However, as AVP is classified as IDP, it is assumed that it exists in an ensemble of conformations. This idea was validated with the ensemble model (see Section 4.7.2). Figure 4.19 shows that the best agreement with experimental chemical shift values is for the REMD *Twisted Saddle* ensemble in which the most populated structures are *Saddle* and *Twisted Saddle*, followed by *Open* and *Clinched Open* which adopt very similar populations. However, all the computationally derived ensembles show improved agreement between the calculated and experimental chemical shifts, over and above that for any single ensemble cluster conformation. This suggests that AVP adopts flexible conformational ensemble with no single preferred structure state which is in agreement with structural data in the literature. Moreover, the ensemble model built from optimised population ratios also confirms that the simulation population ratios yield meaningful results.

Since AVP binds to the same receptor as Oxytocin, the conformational flexibilty of that peptide was also examined. The simulation data and results are given in the next chapter, after which the conformational ensembles are compared for both peptides.

Chapter 5

Oxytocin

Oxytocin (OXT, OT) is another example of the cyclic peptide hormone with structural motif of a 6 membered ring with C terminal tail. It shares the same ring sequence with AVP *(Cys^1, Tyr^2, Ile^3, Gln^4, Asn^5, Cys^6)* but they differ in a third position residues; OT has *Ile* instead of *Phe*. The Oxytocin C terminal tail consists of 3 residues Pro^7, Leu^8, Gly^9 capped with NH_2 group (Figure 5.1). A tail part is also different from AVP in one residue, positively charged and hydrophilic Arg^8 is replaced with hydrophobic Leu^8. The amidated C terminal is a natural form of OT [183], and the total charge of +1 comes from the protonated N-terminus Cys^1 (NH_2).

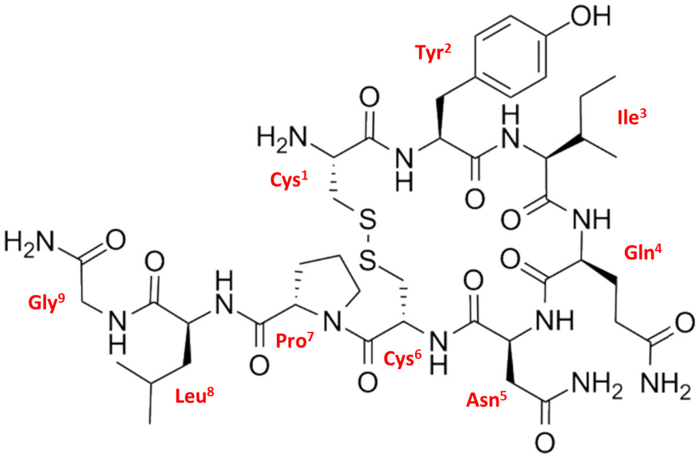


Figure 5.1: Oxytocin is a cyclic peptide made of six ring residues (Cys^1, Tyr^2, Ile^3, Gln^4, Asn^5, Cys^6), and three tail residues *(Pro^7, Leu^8, Gly^9)* capped with NH_2 group.

5.1 Experimental data

Oxytocin peptide was already known in 1953 when its sequence was determined [183], while the 3D structure of its analog, deamino-oxytocin was determined soon after, in 1964 to 1966 [184–186]. Deamino-oxytocin (dOT) has $Mpa - Tyr^2 - Ile^3 - Gln^4 - Asn^5 - Cys^6 - Pro^7 - Leu^8 - Gly^9$ sequence, and it differs from OT only in the N-terminal amino group, so it is often considered as an OT model structure. The crystallographic dOT structure was refined in two studies [187, 188], resulting in two PDB structures with IDs, 1XY1 and 1XY2 (Figure 5.2 (b)). The OT structure was crystallographically determined only in complex with its carrier protein Neurophysin published in 1996 (PDB ID: 1NPO) [189] (Figure 5.2 (a)). **X-ray structures** were described as adopting following conformations:

- Two crystal dOT structures (PDB IDs: 1XY1, 1XY2) are characterized by β-turn II centred at residues 3,4 and occupying $Tyr^2O - Asn^5H$ and $Asn^5O - Tyr^2H$ hydrogen bonds in the ring. The tail is described with 7,8 β-turn III and $Cys^6O - Gly^9H$.

- Oxytocin structure bound to NP (PDB ID: 1NPO) is characterised with β-turn centred at residues 3,4 as well, but as type III. There was no report of hydrogen bonds. The tail is described as crystallising in two forms, *folded* and *extended*. The folded conformation exhibits a 7,8 β-turn, while the extended conformation is characterised by a $Pro^7O - Gly^9H_{NH_2}$ hydrogen bond.

Oxytocin was also extensively analysed with **NMR**. Two groups reported a folded-like ring conformations in water. Ohno et al. [190] characterised ring conformation with 3,4 β-turn stabilised by the two $Tyr^2O - Asn^5H$ and $Asn^5O - Tyr^2H$ ring hydrogen bonds. The tail was described with 7,8 β-turn, and the $Cys^6O - Gly^9H$ hydrogen bond between ring and tail residues.

Another experimental group (Koehbach et al. [191]) characterised OT ring conformation with 3,4 β-turn without reporting on any hydrogen bond, or the β-turns in the C terminal tail [191]. This experimentally averaged ensemble of 50 structures can also be found as PDB code (PDB ID:2MGO).

NMR experiments in DMSO report again a β-turn centred at residues Ile^3, Gln^4 with $Tyr^2O - Asn^5H$, $Tyr^2O - Cys^6H$, and/or $Asn^5O - Tyr^2H$ hydrogen bonds [192–194].

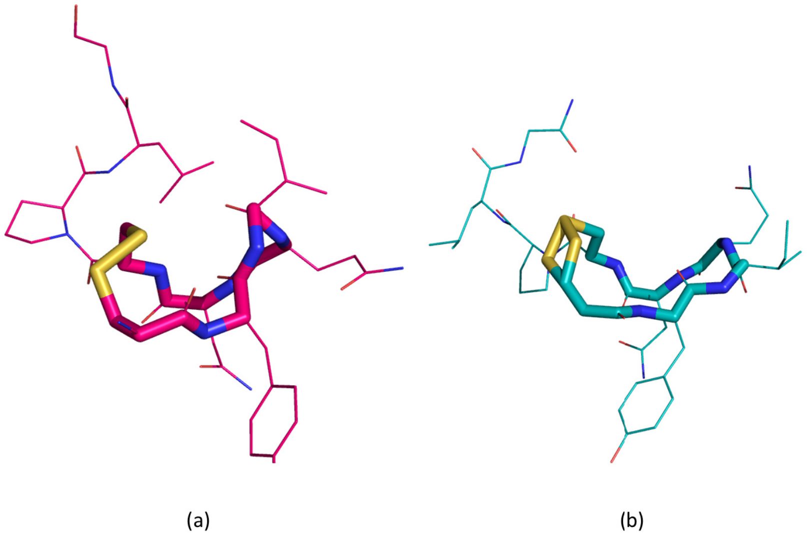


Figure 5.2: (a) Oxytocin crystal structure bound to Neurophysin (PDB: 1NPO). The ring structure is in the *Saddle* conformation. (b) Deamino-oxytocin (dOT) crystal structure (PDB: 1XY1). The ring structure resembles the *Twisted Saddle* conformation.

To summarise, the ring part of the OT seems to be more conformationally rigid than AVP since almost all experimental studies describe it with a rather *folded* ring conformation with 3,4 β-turn stabilised by the $Tyr^2O - Asn^5H$, $Tyr^2O - Cys^6H$ and $Asn^5O - Tyr^2H$ hydrogen bonds.

5.2 Computational data

Oxytocin is often studied together with Vasopressin because of their structural similarity and biological importance. The first of these studies explored conformational ensemble of Oxytocin using a combination of Monte Carlo and MD method for 400 ps. Suggested conformations have β-turns centred at residues 2,3 and 3,4 [172].

The next study was with reservoir REMD in the ff99SB-ILDN force field and TIP4P-Ew water model [174]. In this study, they were comparing the conformational ensembles between AVP and OT. The suggested ring conformation was described as *canonical* with $Tyr^2O - Asn^5H$ and $Tyr^2O - Cys^6H$ hydrogen bonds. However, compared to AVP, Oxytocin adopted a higher percentage of the *extended* tail conformation relative to the *compact* tail subpopulation characterized by either $Cys^6O - Gly^9H$ or $Pro^7O - Gly^9H$ hydrogen bonds.

Finally, the OT conformational ensemble was also explored using normal Molecular Dynamics simulation in explicit water for 50 μs in total over four simulation repeats [182]. The simulation detected the same two main conformational states as in AVP, more open and folded-like conformations, with their states named the same as for AVP, *Open, Clinched Open, Saddle* and *Twisted Saddle.* They also reported on a few very low populated variants of the main conformational states but these were considered as transient states. The tail was reported to be in two conformations, *folded* and *extended* in approximately 20:80 ratio.

5.3 Motivation for our work

Oxytocin and Vasopressin are two cyclic peptides which share the same structural motif of tail attached to ring closed by disulphide bridge. They differ only in the third and eighth residues in the sequence. In Oxytocin, Ile^3 is in place of Phe^3 for Vasopressin, while Leu^8 is instead of Arg^8 in the tail part of the peptide.

Both peptides bind to the same GPCRs on the cell membranes [195, 196], but with different affinities [197]. It is not clear whether it is the different ring conformation that affects affinity or the interactions with the tail.

Moreover, the experimental data reviewed here suggests that OT adopts only folded structures, but similar was claimed for AVP, especially from NMR experiments. Here we explore the conformational ensemble of OT as was done for AVP, and then compare the two conformational ensembles. To do so, Replica Exchange with Solute Tempering (REST) was employed which enhances sampling by softening interactions across a number of replicas keeping the solvent as per the lowest temperature replica (see Section 3.4.4.2 for details).

5.4 REST simulation details

The Solute Tempering simulations were run in Gromacs software using the Amber14 force field. The method was implemented in Gromacs with the Plumed patch [155]. The peptide was simulated in a TIP3P water model [177] containing 1696 and 1961 water molecules, for *Saddle* and *Open* starting conformation simulations, respectively. Furthermore, the system was neutralised with a Cl^- counterion. Particle Mesh Ewald [126] was used for the long-range interactions using a 10 Å cutoff. Bonds involving hydrogen were constrained using the SHAKE algorithm [121] with a tolerance of 0.00001 Å. REST simulations were performed in the NVT ensemble using a Langevin thermostat for the temperature coupling with a collision frequency of 1 ps^{-1}.

The simulations were run for 300 ns using 12 replicas in the effective temperature range 298 K - 900 K. The replicas were geometrically distributed to give the acceptance ratio between 20 and 35 %. In the REST method compared to REMD, the higher temperature range is possible because in REST all temperatures, except the lowest, are taken as pseudo temperatures only, and used to scale the interactions to speed up the sampling, and are not physically meaningful.

5.4.1 The efficiency of REST simulations

The efficiency of Replica Exchange simulations is usually checked by looking at the replica random walk between highest and lowest temperatures, and calculating the acceptance probability between the replicas which ensures that the neighbouring replicas are overlapping enough to allow for efficient configuration exchanges.

The same was done for the REST simulations. Figure 5.3 (a) shows that the lowest replica trajectory visited the complete temperature space, while the desired acceptance probability of 20 - 35 % was achieved (Figure 5.3 (b)).

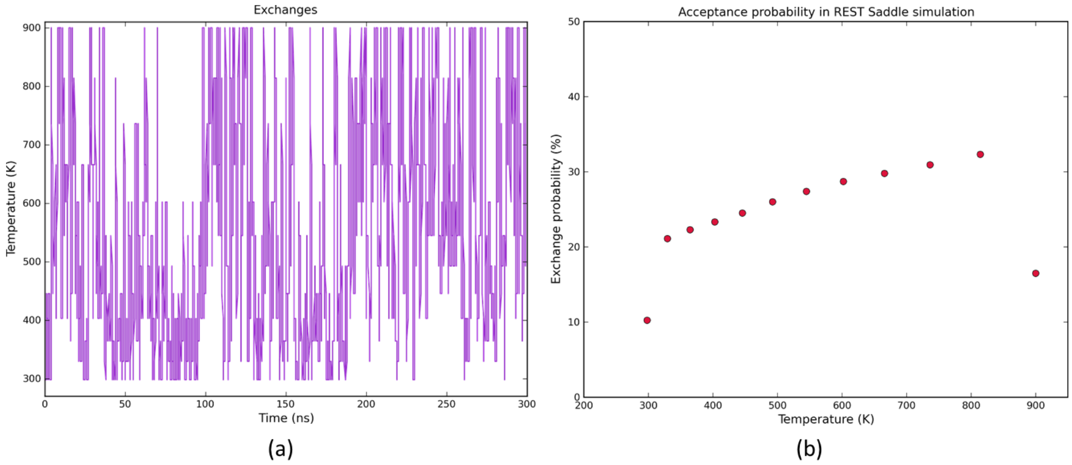


Figure 5.3: (a) Random walk of the lowest temperature replica from the REST *Open* simulation. (b) Acceptance probability between neighbouring replicas was calculated from the REST *Saddle* simulation.

5.5 REST simulation results

Using the REST method, two simulations were run; one starting with *Open*, and another one starting with the *Saddle* conformation. The *Open* conformation was obtained from a high-temperature (800 K) short-scale MD simulation by our collaborators from the University of Portsmouth, and then given to us. The *Saddle* conformation corresponds to the crystal structure co-crystallised in complex with Neurophysin (Section 5.1).

The simulations were performed for 300 ns in total per replica. However, the initial 100 ns of each simulation was taken as equilibration time and was not included in the final analysis. The peptides were analysed in terms of β-turn population, hydrogen bond population and cluster state diversity, just as was done for AVP. The population of the *cis* amide bond next to Pro^7 was also checked.

5.5.1 β-turn and hydrogen bond populations

Following the same approach as for AVP peptide, first the population of β-turns and hydrogen bonds in the simulation trajectories were analysed. The population of different β-turns is given in Table 5.1.

	2,3 type I	2,3 type II	3,4 type I	3,4 type II	4,5 type I	4,5 type II	7,8 type I	7,8 type II
REST Open	0.01	0.95	56.48	4.33	34.69	0.66	5.28	9.67
REST Saddle	1.22	0.83	49.7	12.54	30.57	1.20	5.09	9.57

Table 5.1: β-turn type populations from the two REST simulations

As can be seen from the Table 5.1, OT adopts certain β-**turns** between the ring residues, in particular 3,4 and 4,5 centred turns. The 3,4 β-turn is highly populated, 55 to 70 %, followed by the 4,5 β-turn between 30 to 35 %.

The β-turn population was also checked for the tail residues as both crystallographic structures reported the appearance of 7,8 β-turn. In our simulations, this turn appeared with the approximate population of 15 %.

Compared with the population of AVP β-turns (Table 4.1), then it can be noticed that both peptides adopt the same β-turn types, but OT has higher percentages for both β-turns. This could imply that OT is more conformationally constrained than AVP.

Next, the simulations were analysed with a **hydrogen bond analysis**, because it is known from experimental data that the OT structure is stabilised with certain hydrogen bonds (see Section 5.1). Based on the hydrogen bond definition introduced in Section 1.2.3, the populations of certain hydrogen bonds in the trajectory were analysed.

O - - H	**REST Open**	**REST Saddle**
$Cys^1 - Gln^4$	0.81	0.69
$Tyr^2 - Gln^4$	4.56	4.24
$Tyr^2 - Asn^5$	80.75	78.43
$Tyr^2 - Cys^6$	3.29	4.31
$Ile^3 - Asn^5$	0.98	1.82
$Ile^3 - Cys^6$	1.25	2.75
$Gln^4 - Cys^6$	0.97	0.96
$Asn^5 - Tyr^2$	0.05	0.14
$Cys^6 - Gly^9$	0.66	10.88

Table 5.2: Hydrogen bond population in the two REST simulations named as *Open* and *Saddle* due to the conformation of the starting structures.

Table 5.2 shows that the OT structure was stabilised with a $Tyr^2O - Asn^5H$ hydrogen bond between two ring residues during the largest period of the analysed 100 ns to 300 ns simulation time (almost 80 % of time). Two other hydrogen bonds between ring residues appeared during a short period of time (5 %), $Tyr^2O{-}Gln^4H$ and $Tyr^2O - Cys^6H$. The hydrogen bond between the ring residue Cys^6O and the tail residue Gly^9H appeared during 10 % of the simulation time. Moreover, different hydrogen bond populations across two simulation repeats show very good agreement (Table 5.2).

The high population of the $Tyr^2O - Asn^5H$ hydrogen bond is in agreement with β-turn populations, suggesting that OT prefers the experimentally reported folded-like ring conformation with the $Tyr^2O - Asn^5H$ intracyclic bond.

5.5.2 *cis/trans* proline peptide bond

Since a proline residue is present in OT sequence, the *cis* population of the amide bond next to proline was also analysed. The *cis*-proline bond was defined as adopting range of +/- 60° from the mean value of 0° around the $N_{cys} - CA_{cys} -$

$C_{pro} - N_{pro}$ bond. The results are given in Table 5.3, in which it can be seen that *cis*-proline amide bond conformation is presented in both simulation ensembles.

The experimental population of this amide bond was reported to be approximately 10 % in one study [179].

amide conformation	REST Open	REST Saddle
cis	3.94	5.51
trans	96.06	94.49

Table 5.3: The populations of the *cis/trans* amide bonds during the REST simulations.

5.5.3 Cluster population

Furthermore, OT conformational ensemble diversity was also tested with the torsion based clustering software Dash, following the same approach as for AVP. The cluster population analysis was performed on the lowest replica trajectory at T=298 K, from 100 ns to 300 ns replica time, where the first 100 ns were considered as equilibration time and were not included in the final analysis.

The first step taken to analyse trajectories included the extraction of the $(\phi\psi)$ torsion angles for the ring residues Tyr^2 to Cys^6. The torsion angle values were then run with the Dash software to produce the list of several OT Dash states, then utilised to check for the OT conformational diversity with *dashsim* program.

The list of Dash states also contains the mean values with associated SD of the ring torsions. A *dashsim* program compares these Dash state torsion values between themselves, and reports on the similarity between them. Therefore, it allows us to calculate the similarity of Dash states identified for two different peptide conformations. In total, a unique combination of ten torsion angles defines each cluster state.

5.5.3.1 Oxytocin cluster states

From the torsion angle based cluster analysis, OT was recognised to adopt four cluster states, *Open, Clinched Open, Saddle* and *Twisted Saddle* (Figure 5.4, Appendix A). They follow the same naming as AVP because when compared to AVP ensemble members, they have very high values of the circular similarity scores (Table 5.4). The overlapping between the OT and AVP cluster states is also shown in Figure 5.5, where the structures are aligned on the ring backbone residues.

	OT Saddle	OT Tw. Saddle	OT Cl. Open	OT Open
AVP Saddle	0.97	0.60	0.51	0.53
AVP Tw. Saddle	0.62	0.97	0.55	0.41
AVP Cl. Open	0.52	0.56	0.97	0.38
AVP Open	0.53	0.41	0.38	0.97

Table 5.4: The circular similarity between AVP and OT cluster members

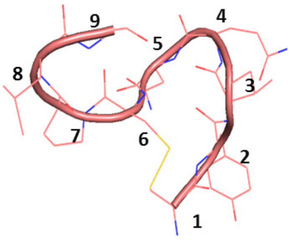
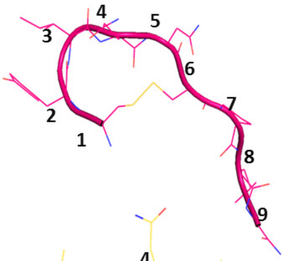
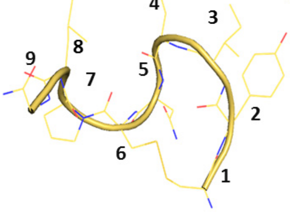
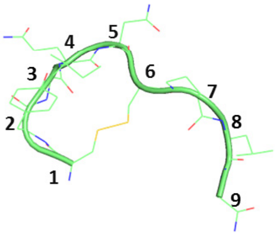
Cluster state	Cartoon representation	Turn type
Saddle		**3,4** and **4,5 β-turns**
Open		No turns
Twisted **Saddle**		**3,4 β-turns**
Clinched **Open**		**4,5 β-turns**

Figure 5.4: The cluster states adopted by Oxytocin with associated turn types.

5.5.3.2 Cluster state time evolution

Having defined the OT cluster ensemble, we also wanted to check for their time evolution. The approach taken is the same as for AVP, described in Figure 4.7.

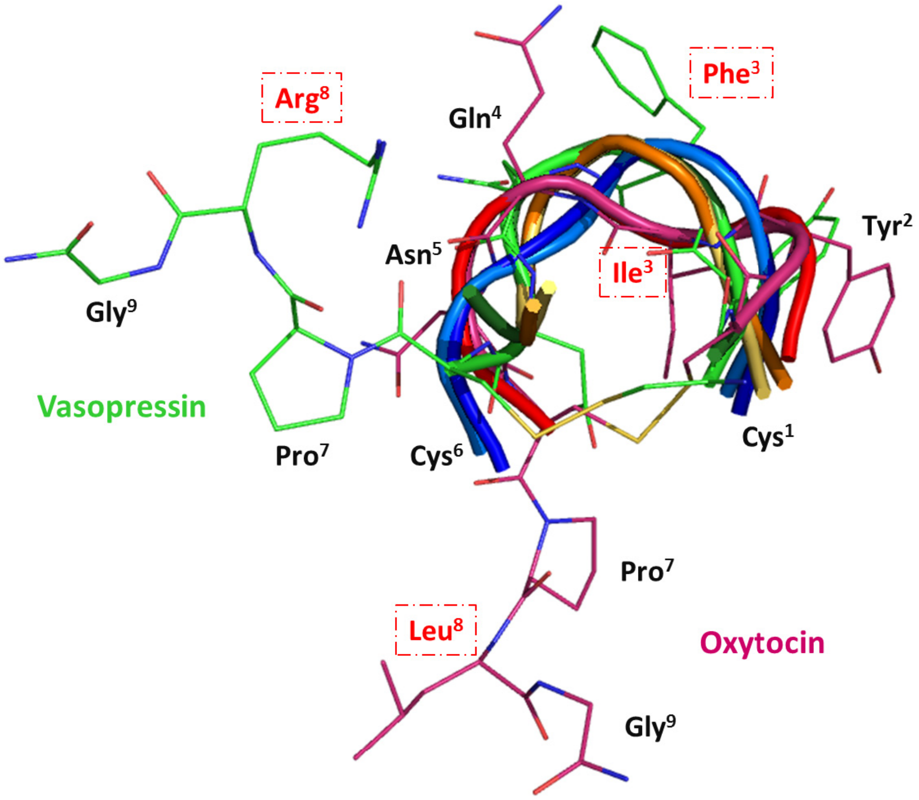


Figure 5.5: The Vasopressin and Oxytocin cluster members aligned based on Cartesian superimposition of backbone atoms on ring part of the structures. In blue are shown the *Open* conformations, while in red *Clinched Open* conformation. *Twisted Saddle* AVP and OT cluster states are shown orange, and *Saddle* is shown in green. The ring is in cartoon representation. The residues which are different between peptides are emphasised in red boxes, and the ones which are the same are given in black. AVP *tail* ($Pro^7 - Arg^8 - Gly^9$) is given in green, and OT *tail* ($Pro^7 - Leu^8 - Gly^9$) is in purple.

The OT cluster state populations across the REST simulations is given in Figure 5.6. It shows that OT prefers the *Saddle* conformation over other OT cluster states. This conformation is populated between 70 and 90 % on average during the simulation time. The second most populated conformational state is *Twisted Saddle*, which is more populated in REST *Saddle* than in the REST *Open* ensembles. Two other cluster states, *Open* and *Clinched Open*, appeared less than 5 % in both ensembles.

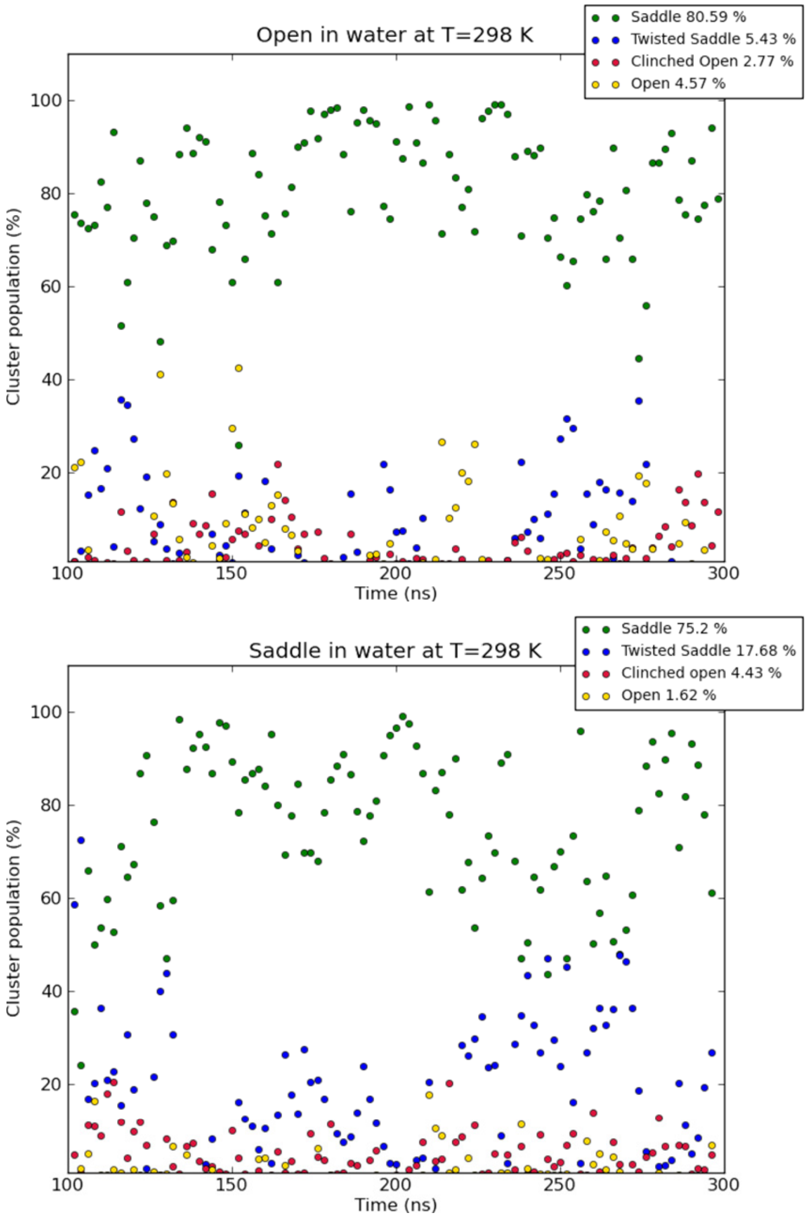


Figure 5.6: The population time evolution of the individual cluster states in two REST simulations of Oxytocin.

5.5.3.3 Comparison between cluster states and crystallographic structures

The obtained OT ensemble cluster states were compared against two crystallographic structures (PDB IDs: 1NPO and 1XY1) in terms of the adopted $\phi\psi$ ring torsion angles. A Dashsim program was used to assign circular similarity between the states.

Two states can be considered similar if the circular similarity between them is higher than 0.65. This number was taken because when the cluster states were

Cluster state	1NPO	1XY1
Saddle	0.78	0.59
Twisted Saddle	0.57	0.80
Clinched Open	0.56	0.58
Open	0.53	0.51

Table 5.5: The OT cluster states compared in terms of circular similarity with two crystallographic structures with PDB IDs: 1NPO and 1XY1.

visually aligned based on Cartesian superimposition of backbone atoms in ring part of the structures, then the ring backbone conformation between two peptide conformations would be similar enough to consider the structures as belonging to the same cluster state.

Table 5.5 shows that 1NPO crystal structure is very similar to the OT *Saddle* cluster state, while 1XY1 crystal structure is very similar to the *Twisted Saddle* OT cluster conformation. This data suggest that the states we observed as the most populated during the simulation times are resembling crystallographic structures. The X-ray determined structures are visualised in Figure 5.2.

5.5.4 The population of *cis*-proline amide in OT cluster states

Table 5.3 reports that in REST simulations the *cis* amide bond associated with the proline residue appeared with a population of approximately 5 %. However, now that the OT cluster ensemble is defined, the selectivity for any of the OT cluster states was tested.

Figure 5.7 shows that *cis* isomer is mostly present in the *Saddle* conformation, while less presented in other conformations. The detailed population of *cis*-proline amide bond in each cluster state is given in Table 5.6.

The *cis*-proline amide population in Table 5.6 does not add up to 100 %. This is because the *cis*-proline amide population is also present in unassigned states of very low population.

The data suggest that the *cis* isomer is not selective for any OT cluster state. Since *Saddle* conformation is the preferred OT conformational state, the *cis* bond was also mostly populated in that cluster. Other cluster states show similar *cis* populations, especially the *Clinched Open* cluster state that seems to adopt con-

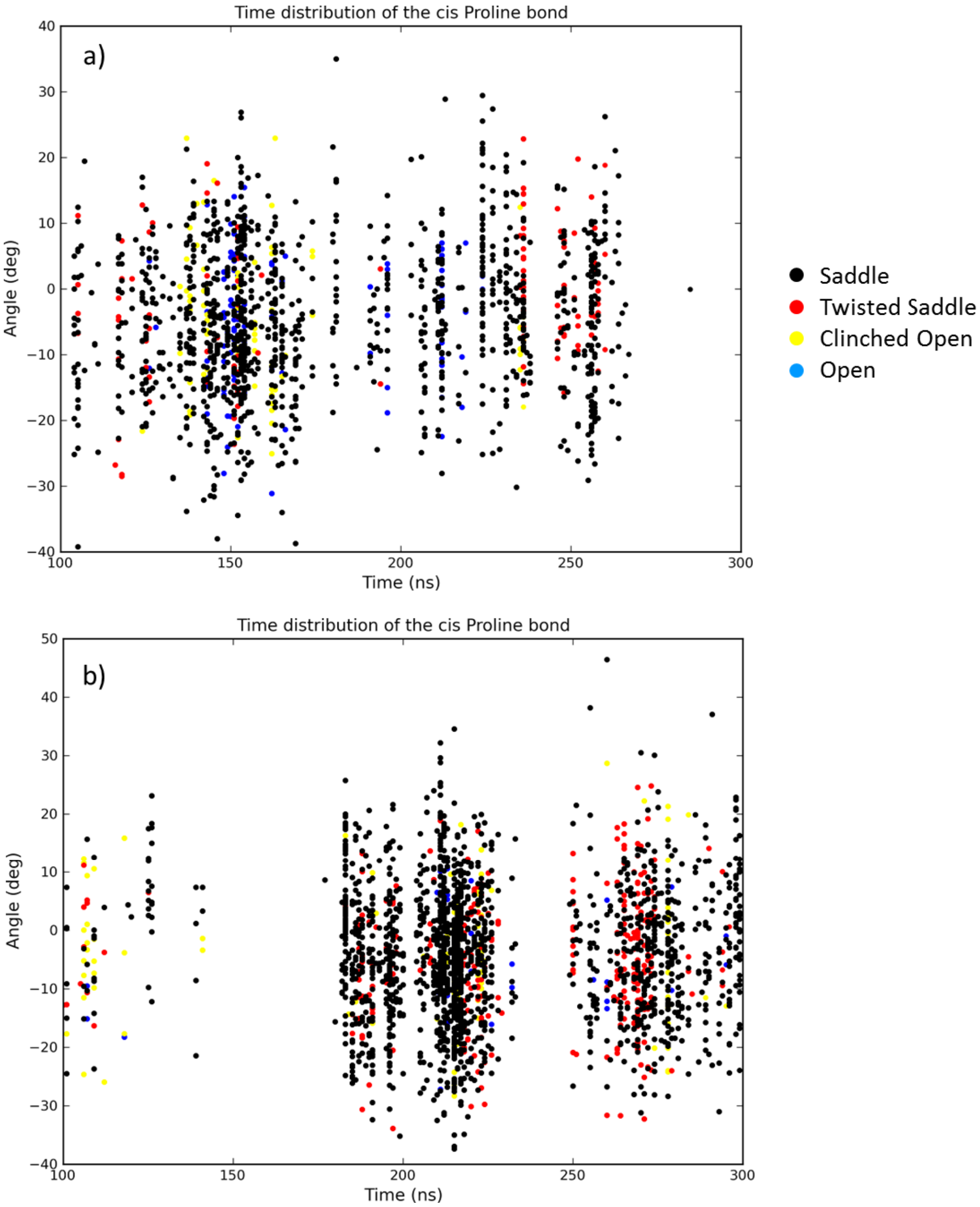


Figure 5.7: A distribution of the *cis* amide bond during simulation time for each OT cluster state in a) REST Open and b) REST Saddle simulations. The colour coding is different for each cluster state.

sistent *cis* populations, while for the *Twisted Saddle* and *Open* cluster states it was more simulation specific because there is approximately 5 % difference in the population of *cis*-proline amide across the simulation ensembles for these two OT conformational states (Table 5.6).

Cluster state	REST Open	REST Saddle
Saddle	75.03	78.43
Twisted Saddle	7.61	13.16
Clinched Open	4.95	5.85
Open	6.97	1.54

Table 5.6: The population of *cis* amide bond in each OT cluster state in the two simulation repeats.

5.5.5 Summary

By performing enhanced sampling REST Oxytocin simulation, we sought to obtain converged OT cluster ensembles. The literature review showed that OT prefers folded-like ring conformational state (see Sections 5.1 and 5.2), named as *Saddle* conformation in our work. This result was also confirmed by our work. Figure 5.6 clearly shows that *Saddle*, followed by the *Twisted Saddle* conformation is the stable OT conformation in aqueous solution. The OT preference for rather folded conformation was also confirmed by the β-turn and hydrogen bond populations (Tables 5.1 5.2). This analysis confirmed the presence of 80 % of the $Tyr^2O - Asn^5H$ hydrogen bond thought to stabilise folded-like ring conformations (*Saddle* and *Twisted Saddle*). The most populated conformations in the REST ensembles resemble two OT X-ray determined structures as confirmed with circular similarity analysis (see Section 5.5.3.3).

A tail part of the OT peptide was also analysed for the presence of the experimentally reported 7,8 β-turn thought to be responsible for the tail *folded* conformation. In our simulations, it appeared for the 15 % of the simulation time, while hydrogen bond between ring and tail residues $Cys^6O - Gly^9H$ appeared for approximately 10 % of the simulation time. The MD simulations reported that the *folded* tail conformation appeared 10-20 % of the simulation time [182].

Since the OT peptide contains a proline residue which is known to reduce the *cis/trans* energy barrier, the amide bond with which the proline nitrogen is involved was checked for the presence of *cis* isomer. Figure 5.7 shows the distribution of the *cis* amide bond with simulation time. While mostly preferred by the OT *Saddle* cluster state, it appeared in other OT cluster states as well (Table 5.6).

Finally, OT cluster members were identified to be the same as for AVP (Table 5.5 and Figure 5.5). For consistency, both peptides were analysed following the same approach, checking the conformational flexibility of the ring ($\phi\psi$) torsion values. Then, AVP and OT cluster ensembles were compared in terms of torsion values to justify that the ring part of structures adopts the same conformations in both peptides what is visually confirmed in Figure 5.5.

Overall, despite considering OT as an IDP, it is shown here, and in the other papers reviewed, that it prefers more folded conformational states, named as *Saddle* and *Twisted Saddle* in our work. The *Open* and *Clinched Open* cluster states could be considered as transient states for OT.

These simulation results were further examined against the experimental chemical shifts.

5.6 Oxytocin chemical shifts

5.6.1 Experimental chemical shifts

The experimental shifts for OT were measured by 2 groups, Ohno et al. [190] and Koehbach et al. [191]. The Ohno group published the values for 36 proton chemical shifts, while the Koehbach group for 35 proton chemical shifts. The missing proton shift of the Koehbach group belongs to Leu^9 HG atom. The chemical shift values are visually compared in Figure 5.8.

In terms of experimental conditions, the Ohno et al. group recorded chemical shift spectra at a solution pH 6.2 and temperature of 298 K, while Koehabach et al. measured the signal at pH 3.5 and the same temperature of 298 K (Section 5.1).

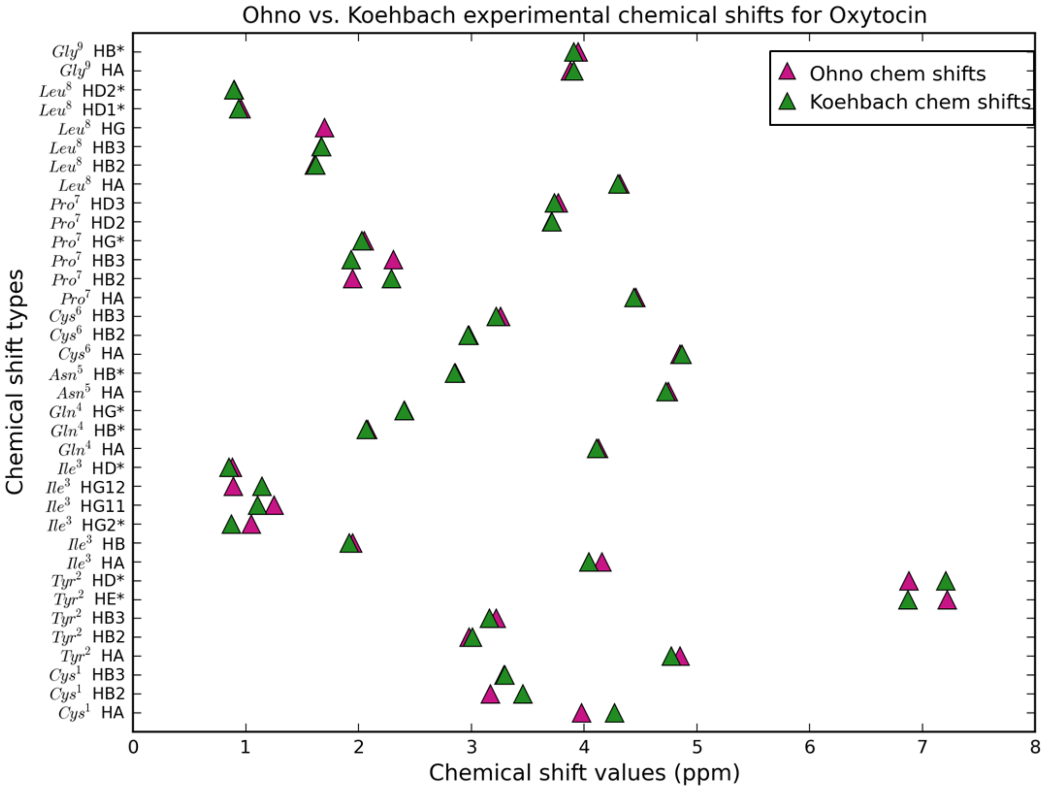


Figure 5.8: The comparison between the values of chemical shifts measured by two groups.

5.6.2 Computational chemical shifts

5.6.2.1 The choice of the representative structures

The representative structures were chosen to fulfil the same conditions as for AVP; to be scattered in approximately equal intervals along the trajectory (Figure 5.10), and to be within the 1 SD of the torsion angle distribution (Figure 5.11). There are 9 representative structures for *Saddle* cluster state, and 10 for other cluster states (*Open, Clinched Open, Twisted Saddle*).

Calculated chemical shifts were compared between the representative structures to see the variance within the chemical shift types. The Tyr^2 HD* and HE*, Pro^7 HD2 and HD3, Leu^8 HD1* and HD2* chemical shift types showed the tightest values across all Oxytocin cluster states (Figure 5.9), while rather wide values of chemical shifts were observed for Gly^9 HA* and HB* chemical shift types, within 0.3 ppm. The Cys^1 HB3 chemical shift type for Twisted Saddle representative structures showed the largest variance, approximately 0.5 ppm.

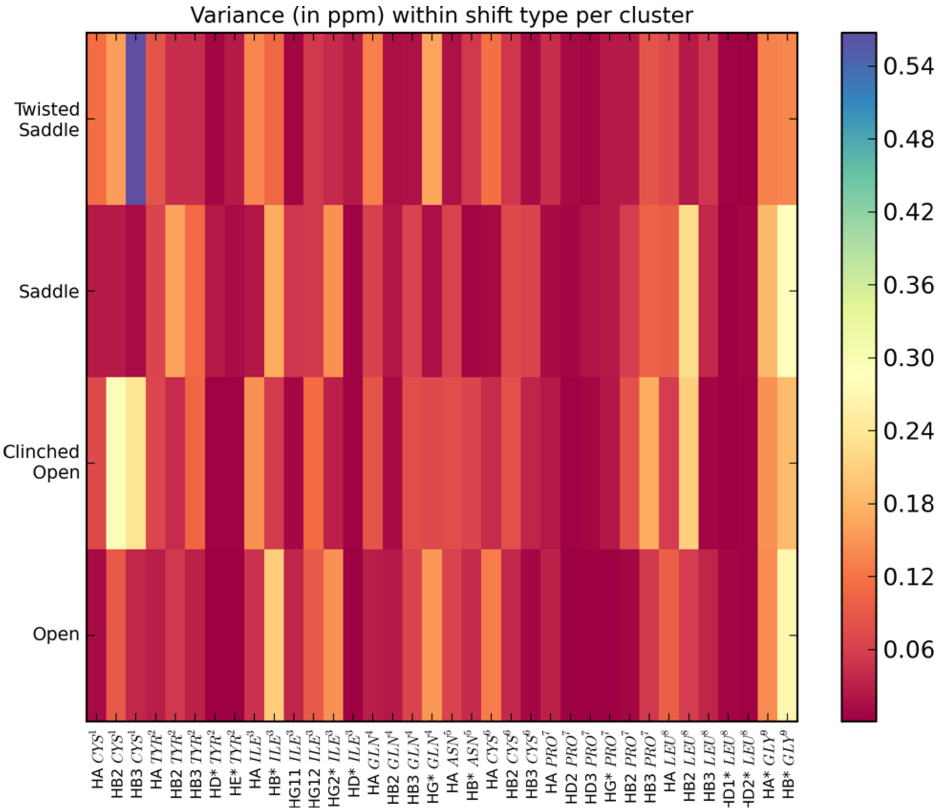


Figure 5.9: The chemical shift type variance for the given representative structures for each Oytocin cluster state.

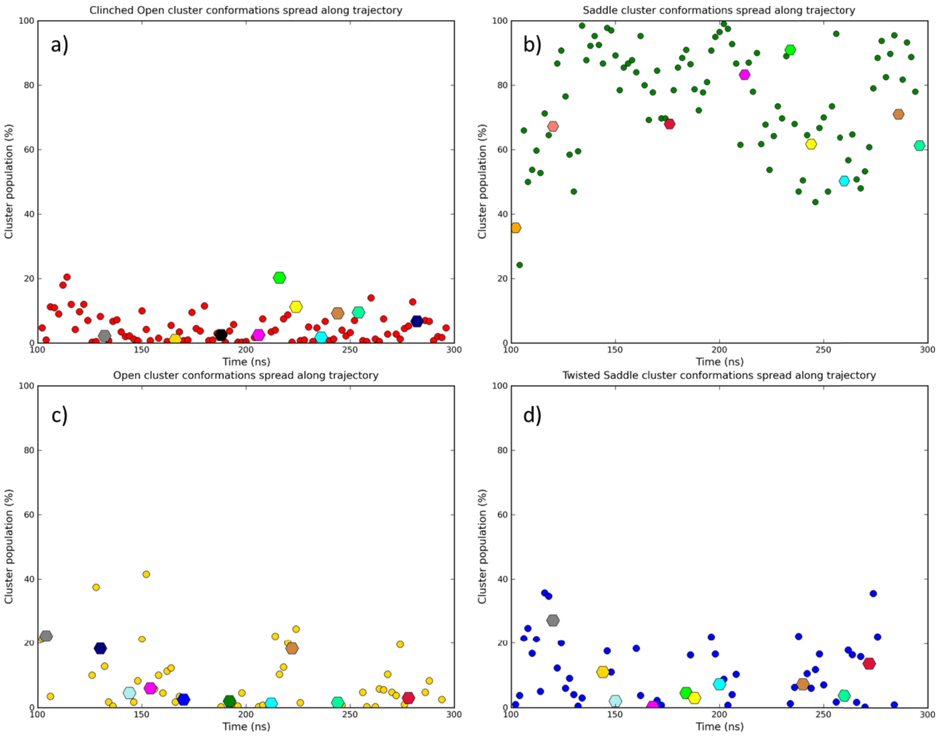


Figure 5.10: The locations of the pulled out representative structures for each cluster state a) Saddle, b) Twisted Saddle c) Clinched Open, d) Open during the simulation time. The emphasised colour dots depict another structure for given cluster state.

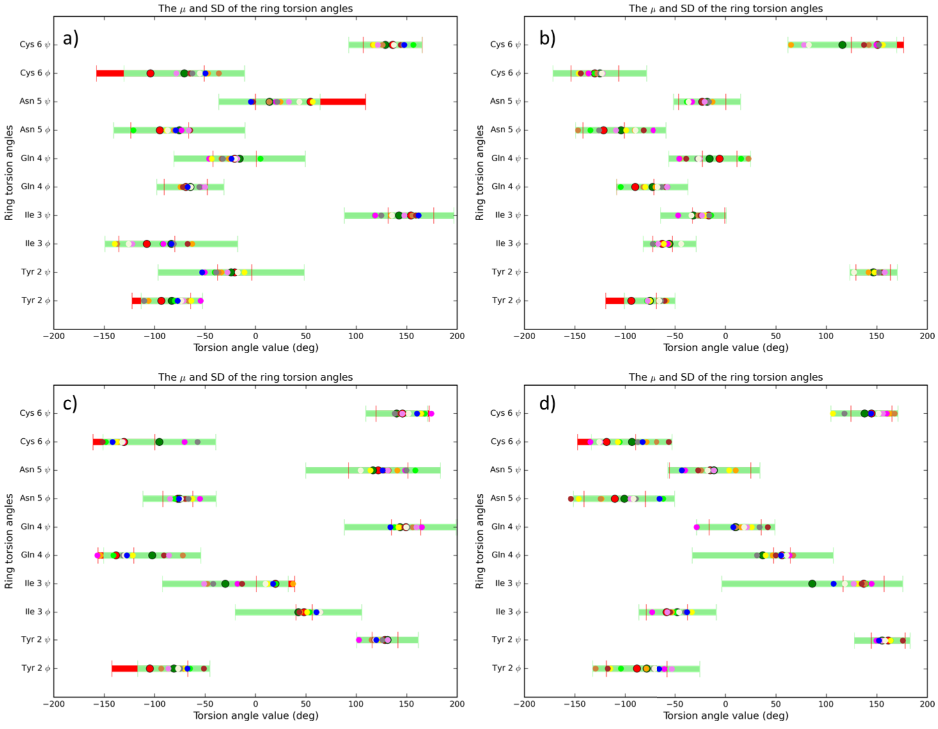


Figure 5.11: The distribution of observed torsion angles for each conformation in each conformational state for a) Open, b) Clinched Open, c) Saddle, d) Twisted Saddle. The red bars are from MD simulations [182], and REST torsion angle distributions are in green. The spots show the dihedral angles for the structures we selected.

5.6.2.2 Chemical shift calculation

After fulfilling the initial requirements, representative structures were optimised and the shielding constants were converted into chemical shifts afterwards using regression equation 4.1 (see Section 4.7.0.2 for details). All the calculations were done with the PCM water model using Gaussian09 software [116] at the B3LYP/6-31G(D) level of theory.

The chemical shift values from optimised structures were then compared with experimental values, with data published from both experimental research groups, in terms of R^2 values. The comparison between each representative structure for each cluster state with experimentally measured chemical shifts is given in Figures 5.12 and 5.13.

By comparing the R^2 range adopted by each individual cluster state, it can be observed that all cluster states adopt very similar R^2 ranges; the *Saddle* and *Twisted Saddle* cluster state conformations have R^2 in the 0.92 to 0.97 range, *Open* structures from 0.91 to 0.95 while *Clinched Open* cluster state structures have tightest range, from 0.93 to 0.96.

Interestingly, the R^2 distribution is rather wide when compared against both sets of experimental chemical shifts, taking values between 0.91 and 0.97.

5.6.3 Bootstrapping of the individual cluster states

Since all the cluster states perform very similar when compared with experimental values, the structures were subsequently bootstrapped to account for the fact that only a few (9 or 10) structures were chosen as a representative of the several hundred frames belonging to a particular cluster. The R^2, MUE and MSE were bootstrapped with 95 % CI. The result of the bootstrapped R^2 distributions are given in Figure 5.14.

The mean R^2 values of the individual cluster states shown in Figure 5.14 are also summarised in Tables 5.7 and 5.8. The values show that no two NMR measurements are reporting the same mean R^2 value for any cluster state but *Saddle*. The biggest R^2 difference was obtained for the *Open* and *Twisted Saddle* cluster states. For the *Open* cluster state the R^2 mean is centred at 0.927 when compared against Ohno chemical shifts, and at 0.937 when compared against Koehbach chemical shifts. A similar difference is observed for the *Twisted Saddle* cluster, with

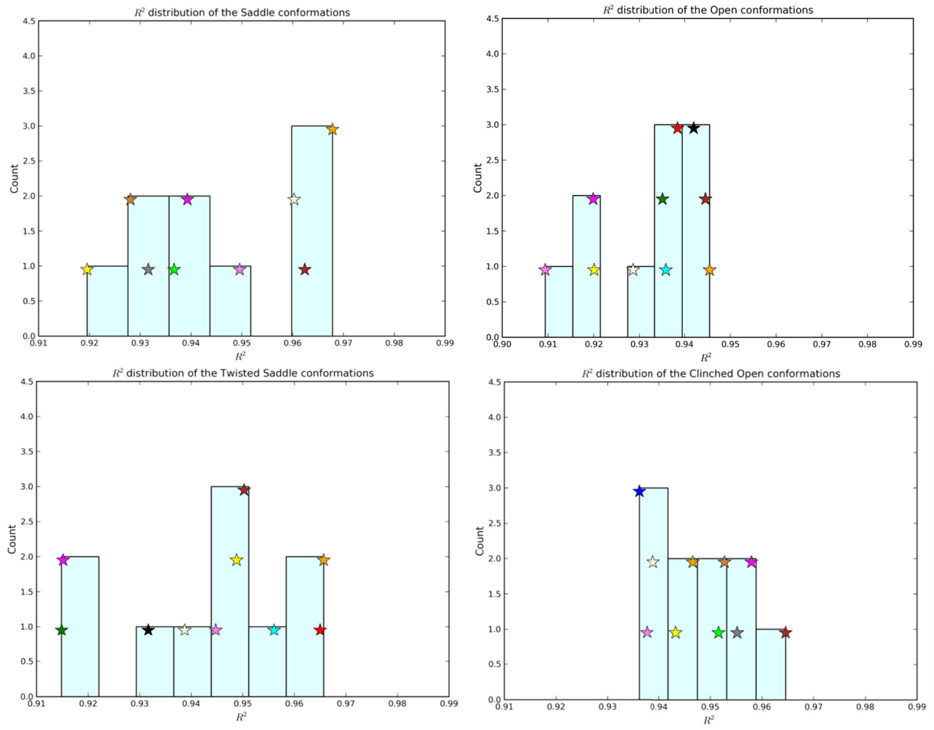


Figure 5.12: The distribution of the R^2 values calculated between experimentally measured and theoretically obtained chemical shifts for each structure belonging to particular Oxytocin cluster. Experimental values were measured by Ohno group [190]. The asterix depicts the R^2 values for each representative structure, and were vertically offset to show their spread within a bar.

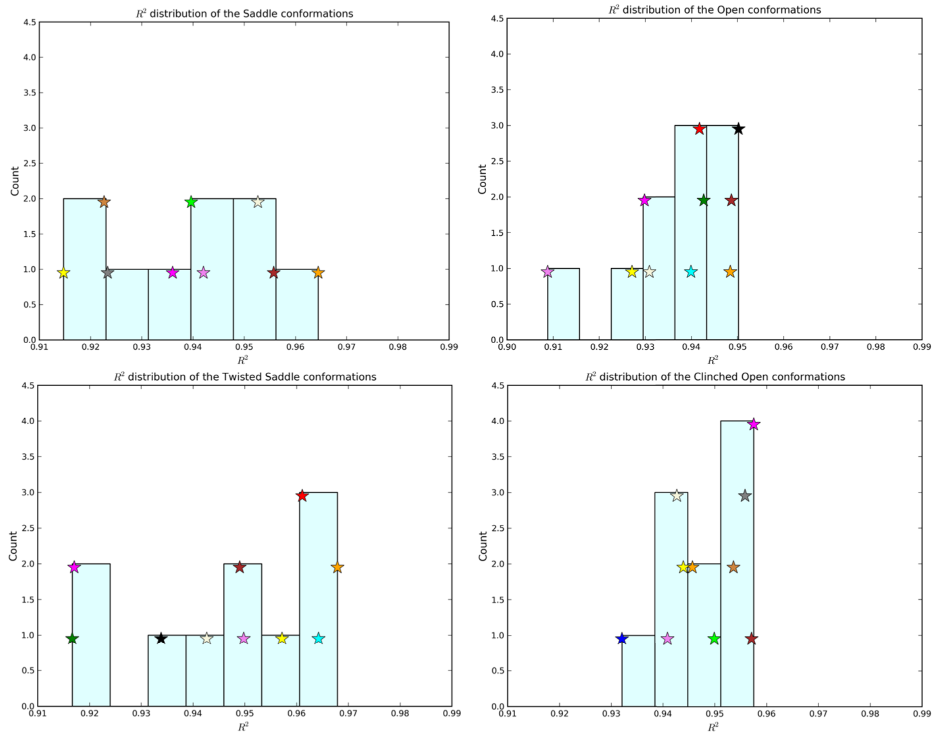


Figure 5.13: Distribution of the R^2 values calculated between experimentally measured and theoretically obtained chemical shifts for each structure belonging to particular Oxytocin cluster. Experimental values were measured by Koehbach group [191]. The asterix depicts the R^2 values for each representative structure, and were vertically offset to show their spread within a bar.

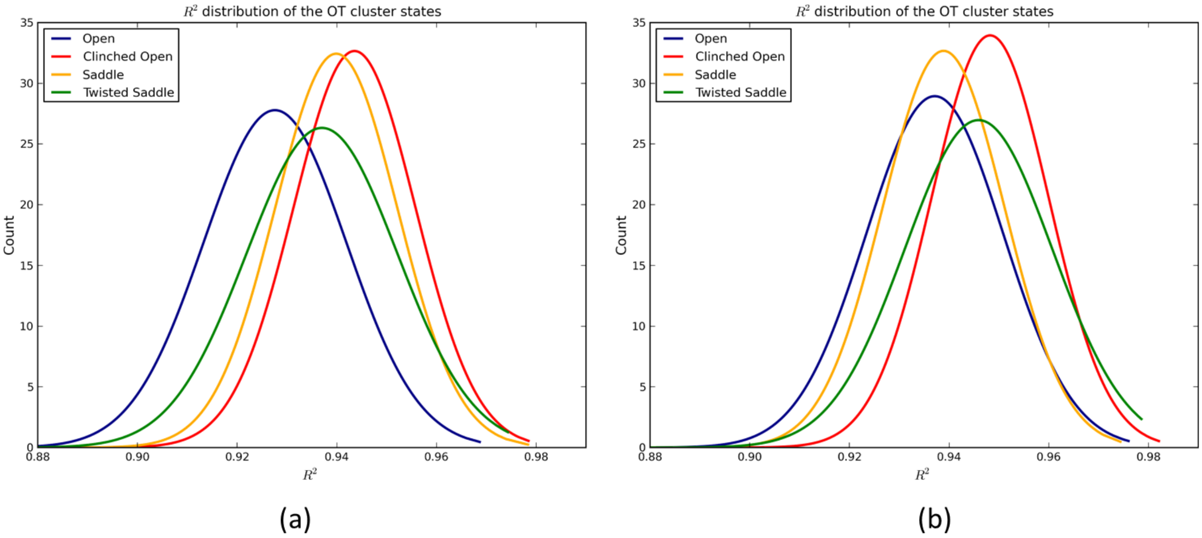


Figure 5.14: Distribution of the R^2 values obtained after bootstrapping calculated chemical shifts from each individual cluster state, and comparing lists with two sets of experimental values, a) Ohno b) Koehbach.

R^2 peaks positioned at 0.937 and 0.946 for Ohno and Koehbach chemical shifts, respectively.

Interestingly, the *Clinched Open* conformation is showing the best agreement with experimental data for both experimental sets, followed by *Saddle* at pH 6.2. Similar MUE values are also reported for *Clinched Open* and *Saddle* cluster states (Tables 5.7 and 5.8).

5.6.3.1 Analysis of the individual chemical shift types

Another part of the analysis of the single Oxytocin cluster states consisted of checking the weight of the individual chemical shift types to the overall R^2 distribution. The chemical shift types were removed from the final analysis in the sequential way, depending on the calculated variance, to check how the correlation coefficient changes with the variance of the individual chemical shift types.

For Oxytocin *Open* cluster state (Figure 5.15 a)) it was observed that the chemical shift types with variance already higher than 0.01 ppm show significant drop in the R^2 value, from 0.94 for the full set of experimental data to 0.85 for data set consisting only of the chemical shift types with variance higher than 0.01 ppm. However, for *Clinched Open* cluster (Figure 5.15 b)), the weight of the chemical shift types with variance lower than 0.005 ppm is smaller than it was

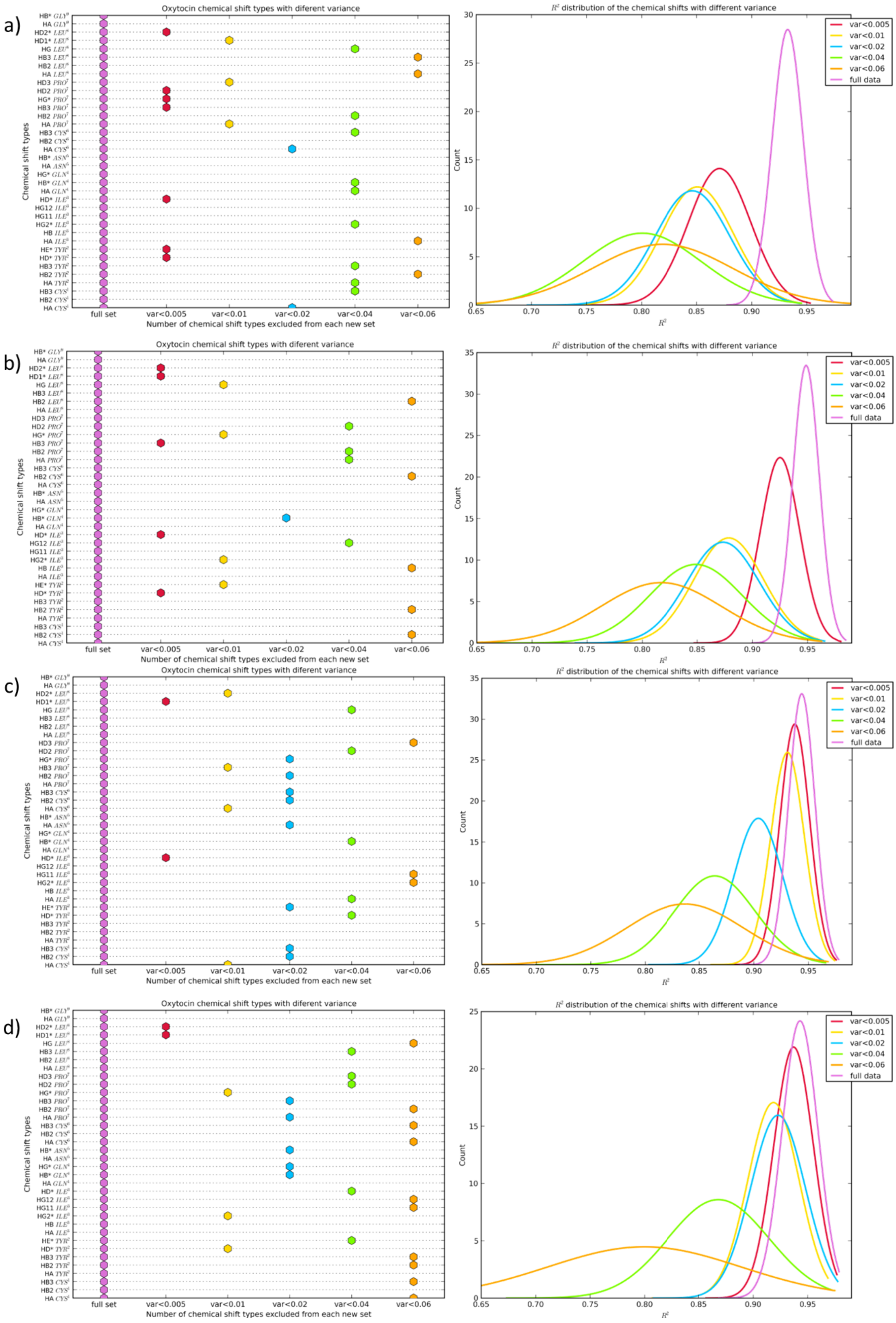


Figure 5.15: The R^2 distributions of the theoretically calculated vs. experimental chemical shifts with different chemical shift data sets depending on the value of their variance (on the left plots). Distributions are plotted separately for a) Open, b) Clinched Open, c) Saddle and d) Twisted Saddle AVP cluster state.

Cluster state	MUE	MSE	R^2
Open	$0.213 < 0.315 < 0.418$	$-0.165 < -0.018 < 0.127$	$0.899 < 0.927 < 0.955$
Clinched Open	$0.216 < 0.284 < 0.352$	$0.003 < 0.113 < 0.223$	$0.919 < 0.943 < 0.967$
Saddle	$0.198 < 0.284 < 0.371$	$-0.035 < 0.089 < 0.214$	$0.915 < 0.939 < 0.964$
Twisted Saddle	$0.217 < 0.291 < 0.363$	$-0.145 < -0.024 < 0.009$	$0.907 < 0.937 < 0.966$

Table 5.7: The bootstrapped values of three statistical measures of similarity, MUE, MSE and R^2 for four OT cluster states when compared with Ohno chemical shifts.

Cluster state	MUE	MSE	R^2
Open	$0.255 < 0.348 < 0.441$	$-0.117 < 0.032 < 0.181$	$0.910 < 0.937 < 0.964$
Clinched Open	$0.197 < 0.267 < 0.336$	$-0.073 < 0.039 < 0.152$	$0.925 < 0.948 < 0.971$
Saddle	$0.206 < 0.273 < 0.340$	$-0.033 < 0.077 < 0.188$	$0.917 < 0.939 < 0.960$
Twisted Saddle	$0.184 < 0.264 < 0.345$	$-0.167 < -0.048 < 0.071$	$0.921 < 0.946 < 0.974$

Table 5.8: The bootstrapped values of three statistical measures of similarity, MUE, MSE and R^2 for four OT cluster states when compared with Koehbach chemical shift values.

for *Open* cluster state. Here the R^2 value dropped only slightly, while for *Saddle* and *Twisted Saddle* cluster states (Figure 5.15 c), d)), the significant decline was observed for chemical shift types with variance higher than 0.04 ppm, which implies that the weight of these chemical shift types is largest to the final R^2 distribution.

5.6.4 Ensemble model

The final step in analysing the chemical shift data is by validating it against the ensemble model. The idea of the ensemble model was introduced in the chapter reporting the AVP peptide results (Figure 4.7.2). The protocol includes weighting each representative structure for each cluster member by the associated cluster state population and taking the weighted sum as a unique set of chemical shifts, to be compared against experimental values.

Figure 5.16 gives the results together with the OT REST simulation populations. From the Figure, it can be seen that the REST ensemble derived by using the *Saddle* as a starting structure, with a high population of *Saddle* cluster state followed by that of the *Twisted Saddle* conformation shows better R^2 with experiment than the ensemble derived using the Open starting structure. This data is in agreement with the two crystallographic OT structures (Figure 5.2) which correspond to our *Saddle* and *Twisted Saddle* cluster states (Table 5.5).

We can also compare the ensemble R^2 mean values between the two experimental sets. It shows that both sets of experimental data are giving very similar mean R^2 values, so both sets can be taken as valid to compare with computational data.

Next, the ensemble R^2 mean values were also compared against the R^2 mean values of the individual cluster states (Table 5.9). Overall, Koehbach experimental shifts are giving higher R^2 values in all cases (individual cluster states and ensembles). This could be due the experimental conditions because these data are measured at lower solution pH.

For the Ohno chemical shifts measured at pH 6.2, the best performing single cluster state is *Clinched Open* (R^2=0.943), followed by *Saddle* (R^2=0.939). However, these values are still lower than for the ensemble models which are giving the values of 0.949 and 0.955 for REST Open and REST Saddle simulations, respectively.

Finally, these data also suggest that the ensemble model is more appropriate to describe the conformational cluster flexibility of cyclic hormone peptides.

Experimental group	Open	Saddle	Cl. Open	Tw. Saddle	REST Open	REST Saddle
Ohno	0.927	0.939	0.943	0.936	0.949	0.955
Koehbach	0.937	0.939	0.948	0.946	0.950	0.957

Table 5.9: The comparison between bootstrapped R^2 values between individual cluster states vs. ensemble model for the two REST simulations (REST Open and REST Saddle), for both sets of experimental chemical shifts.

5.6.4.1 Optimal cluster population ratios

We also wanted to check the cluster population ratios that lead to the optimal prediction of the experimental chemical shifts. Figure 5.17 gives results of such calculation for Ohno (a), and Koehabach (b) experimental chemical shift data.

Population ratios given in Figure 5.17 show that different combinations of cluster state populations give the same result. The simulation cluster populations show that *Saddle* is the most populated conformational state followed by *Twisted Saddle* which yield to the R^2 value of 0.96 (Table 5.9), while population ratios given in Figure 5.17 follow R^2 value higher than 0.985. These results show that different population ratios between the observed Oxytocin cluster states can give very good agreement with experimental data. However, the population ratios with the best agreement of the ensemble model with experimental data show that all identified

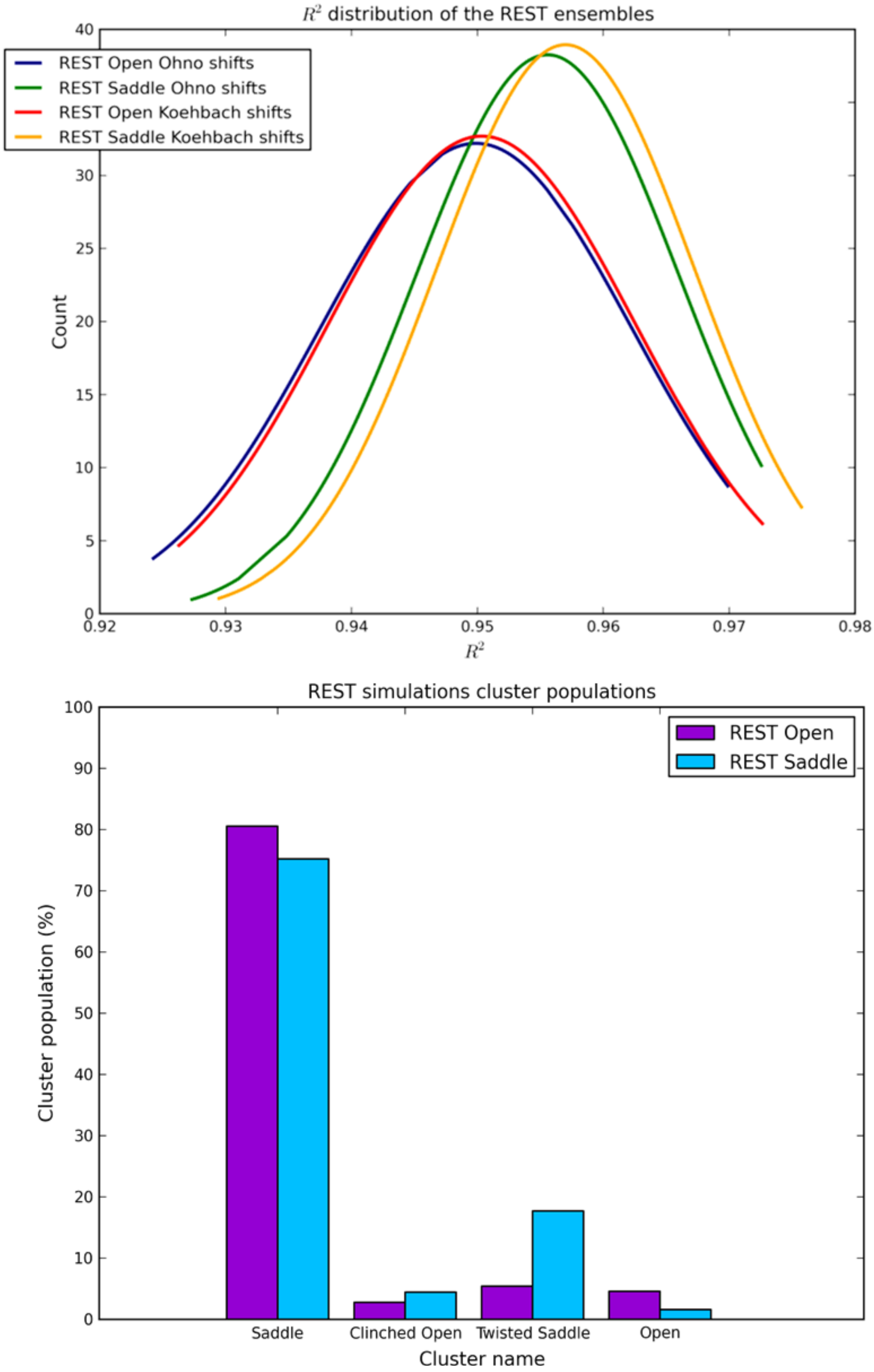


Figure 5.16: The upper part of the picture shows the bootstrapped R^2 distribution of the ensemble model, while the plot below shows the populations of each cluster state in each of the two simulation repeats.

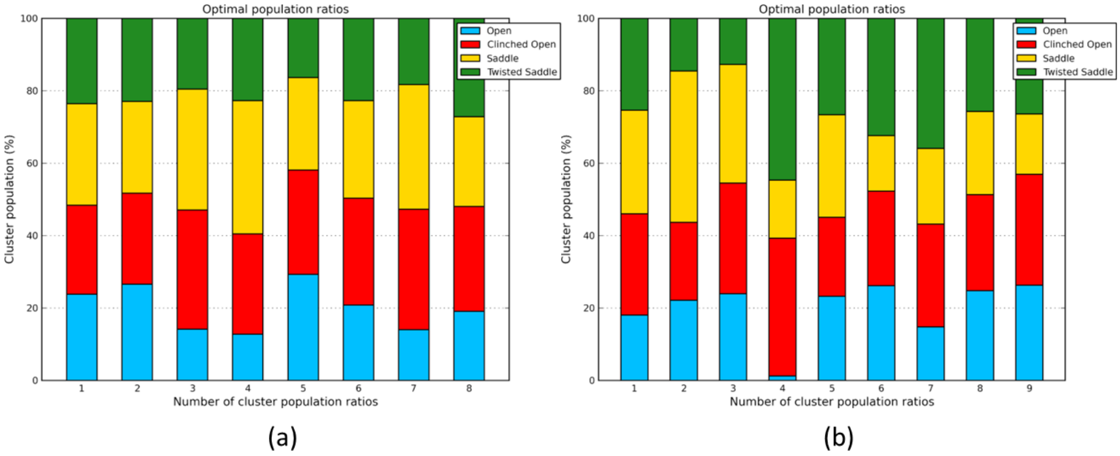


Figure 5.17: The optimal Oxytocin cluster states population ratios that have the best agreement with experimental data from (a) Ohno, (b) Koehbach.

cluster states are similarly populated contrary to simulation results where *Saddle* is the dominant state. This is observed for both sets of experimental data.

5.7 Conclusions

In this chapter, the conformation and dynamics of the cyclic peptide hormone Oxytocin in solution was explored with the combination of enhanced sampling REST simulations and chemical shift calculation. The simulation data suggest that OT prefers a folded-like *Saddle* conformational state which had a high population during the simulation. The REST simulation also revealed that OT can be found in three other minor populated cluster states referred to as *Open, Clinched Open* and *Twisted Saddle.* The states were classified according to the conformation of the ring part of the structure, following the same approach taken in already published data and the AVP analysis given in the previous chapter.

The tail conformation was described as *folded* or *extended* depending on the population of the secondary 7,8 β-turn motif and $Cys^6O - Gly^9H$ hydrogen bond. In our simulations, the population of this beta turn and hydrogen bond was between 10 and 15 %. The MD simulations reported a similar percentage, while the 1NPO crystal structure was co-crystallised with the *folded* tail conformation. Overall, the OT cluster state populations show converged conformational pattern.

The simulation data validation against experimental ^{1}H chemical shifts revealed that the OT prefers the highly populated *Saddle* and less populated *Twisted Saddle* conformational states. The ensemble data derived from the simulation ensembles show higher bootstrapped R^2 values than the individual conformational states. The ensemble model result is supported by crystal OT structures which were also found to adopt these two states, although the *Twisted Saddle* conformation was crystallised for dOT which is an OT model structure.

Since AVP and OT were found to adopt the same cluster states, the next chapter discusses the conformational similarity between AVP and OT.

Chapter 6

Comparing AVP and OT cluster states

Table 5.4 contains the circular similarity scores between AVP and OT cluster states, and it shows that they adopt overlapping cluster ensemble states (Figure 5.5). Their conformational ensembles were explored using two enhanced sampling methods, REMD and REST, which enhance sampling by taking advantage of the high temperatures (REMD) or scaling the potential energy functions (REST). In this chapter, the comparison between the AVP and OT conformational ensembles will be given, together with the overview of their binding affinities to receptors, but first the enhanced sampling methods used will be discussed in terms of computational cost.

6.1 Simulation computational cost

The AVP conformational ensemble was obtained using the REMD method which used 80 replicas to simulate the system for 300 ns per replica. This led to the total of 24 μs of simulation for each repeat. The conformational ensemble of Oxytocin was obtained using another Replica Exchage method in which only the simulated peptide is at different effective temperatures keeping the waters at room temperature; this results in fewer replicas being needed. The REST simulation was run for 300 ns per replica, which together with the 12 replicas used gives 3.6 μs simulation time for each repeat.

Comparing these two methods, there is a clear advantage of REST over REMD. To achieve equilibrium sampling, REMD requires a large number of replicas (80), while for the system of the same size, REST required far fewer (12). This makes REMD computationally expensive because of the need for a highly parallel computational resource. Beside being computationally expensive, it is also time consuming because the large number of replicas are required to exchange at a set time which extends the real time running of the method. Furthermore, Figure 5.3 shows that almost the same acceptance probability is obtained with REST with the smaller number of replicas than with REMD (Figure 4.5). The efficiency of methods to perform a free walk in temperature space was also achieved (Figures 4.5 and 5.3).

In terms of convergence of the conformational sampling, it can be observed in Figures 4.6 and 5.6 that both peptides achieved converged conformational ensembles during the 300 ns of the simulation time per replica. The overall populations of the AVP and OT cluster states show very good agreement across the all simulation repeats.

6.2 Comparison between conformational ensembles

After it was established that the simulations ran properly and efficiently for both peptides, their conformational ensembles were examined.

Figure 5.5 and Table 5.4 shows that AVP and OT are adopting overlapping conformational ensembles, but the ratio of the populations of the individual cluster states differ. Although the folded-like *Saddle* state is the preferred cluster state for both peptides, in AVP (Figure 4.8) it is populated less than in OT, approximately 40 % compared to 75 % (Figure 5.6). Moreover, AVP seems to be more conformationally flexible with *Twisted Saddle, Clinched Open* and *Open* states being similarly populated (10 - 15 % on average). The AVP *Open* structure resembles one of the AVP crystal structure (PDB ID: 1YF4). In OT, the only significant state beside *Saddle* is *Twisted Saddle* which is in agreement with experimental data (Section 5.1). The OT *Saddle* had 5-fold higher population than OT *Twisted Saddle.*

Next, conformational ensembles for both peptides were validated against experimental data. Chemical shifts are a commonly used NMR observables to check computational populations. Here, the equilibrium model which weights the shift values according to cluster population was tested against chemical shift values of the individual conformations. The data confirmed that the AVP and OT ensemble models have better agreement with experimentally determined chemical shift values in solution compared to the individual cluster states, confirming the idea that OT and AVP exist in an ensemble of conformations, and that the enhanced sampling simulations are able to reproduce these experimental ensemble populations.

6.3 Interaction with receptor

Both peptides are endogenous ligands to different GPCR receptors [195, 196]. There are three different AVP receptor subtypes known, V1aR, V1bR and V2R, where V2R is localised on the renal collecting duct and is part of the AVP mechanism responsible for antidiuretic activity (see Section 4.1).

No crystal structure of vasopressin receptors has been reported to date, but the proposed binding poses and ligand interactions are coming from various mutagenesis data [198]. Receptor binding data for AVP proposed that the aromatic side chains of $Tyr^2 - Phe^3$ [199] are interacting with the V2R transmembrane (TM) helices to activate signal transduction. The peptide tail is suggested to be oriented outside the TM core with Arg^8 interacting with the extracellular loop [200].

Oxytocin is though to interact with the OTR receptor [196] via Tyr^2, Ile^3 and Leu^8 residues [201].

While the AVP receptor V2R discriminates between AVP and OT, with AVP binding with 400-fold higher affinity than OT, AVP was discovered to bind to OTR receptor with similar affinity [197, 202], which may suggest that structural difference between OT and AVP could be associated with this selectivity. While their ensembles show overlapping cluster states, their population ratios differ, possibly suggesting that different ring conformations have different biological roles. An AVP tail Arg^8 residue was though to be a key factor in the receptor recognition interacting with the extracellular loop of receptor [170].

There are several proposed mechanisms of peptide binding to GPCR receptor reviewed in Sections 1.3.2 and 2. Most of them suggest that binding events are probably accompanied with the conformational changes to the peptides. However, there is also an evidence that the *bound* conformation of the IDP peptide is found in solution [203] supporting the hypothesis that the converged IDP conformational ensemble contains the peptide bound conformation. Therefore, AVP and OT conformational states can also be considered as candidates for biologically active conformations.

Chapter 7

Urotensin II peptide

Urotensin II (UII) is a cyclic peptide hormone just like AVP and OT. The cyclic part is connected by a disulphide bridge between two cysteine residue ($Cys^5 - Cys^{10}$). N terminal tail is made of four residues $Glu^1, Thr^2, Pro^3, Asp^4$, while C terminal tail contains only Val^{11}. A total aggregate UII charge is -1.

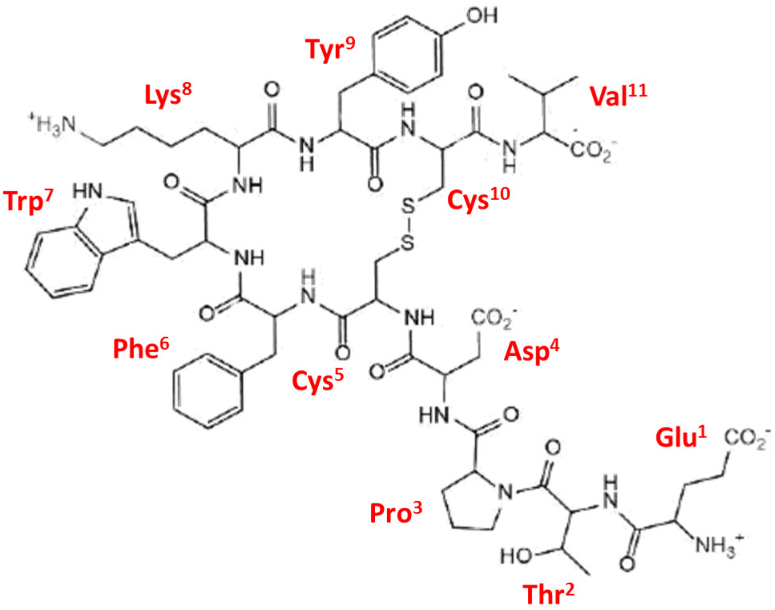


Figure 7.1: Urotensin II is a cyclic peptide made of six ring residues (Cys^5, Phe^6, Trp^7, Lys^8, Tyr^9, Cys^{10}) surrounded by two tails; the N terminal tail contains 4 residues $Glu^1, Thr^2, Pro^3, Asp^4$, while C terminal contains only Val^{11}

7.1 Known structural data

UII was initially found in the urophysis (terminal region of the spinal cord) of teleost fish in 1969 [204]. The human version was identified much later by the three different groups at the same time, in 1999 [205–207]. Since the UII peptide has only recently been identified, there is not much structural data compared to the previously introduced AVP and OT peptides. Moreover, no crystal structure of the UII peptide has been reported to date.

Two NMR studies in water suggest an unstructured ring conformation with no intramolecular hydrogen bond [208], and a widened 7,8,9 γ + 8,9,10 γ ring conformation [209] with the possibility of creating two hydrogen bonds Trp^{7}O-Tyr^{9}NH and Lys^{8}O-Cys^{10}NH. The N terminal tail was described as flexible by both studies.

In DMSO, the ring was described as unstructured, with a possible 3,4 β-turn I in the N terminal tail [210].

The UII structure was also probed in SDS micelles where the ring part of the structure showed *folded* conformational feature, with the β turn type II$'$ centred at residues 7,8 [211].

Overall, only a few NMR studies of the UII peptide report two main UII ring structural features: one describes ring as unstructured, and another as in rather folded conformation, while the N terminal tail was described as flexible.

Regarding the **computational data**, there is only one paper which reported on the rather detailed UII conformational ensemble using a combination of MD and REMD methods [212]. Since we were part of the collaboration which studied this peptide, the REMD part of the results will be given in the results section. On the other hand, the MD simulations, although reported the same UII cluster members as the REMD simulations, the population of the states was dependent on the starting conformation for each MD run. In particular, five MD simulation repeats were performed, each starting with different UII conformational state; four of them run for 5 μs and one MD simulation run for 10 μs. However, the observed UII conformational states in the simulation ensemble were strongly depending on the starting conformation. For example, in the MD simulation started with *Omega I Open* or *Folded I* UII conformational state, only that state was observed for the rest of the 5 μs simulation time. Because of the observed conformational trapping, the UII ensemble was explored using enhanced sampling method.

7.2 Motivation for our work

From Section 7.1, it is clear that conformational data for the UII peptide are rather rare, with no crystal structure obtained yet. Therefore, computational methods could help in getting the complete picture of the UII structural ensemble, in the limit of converged sampling data and force field accuracy.

The UII conformational ensemble was probed with two enhanced sampling methods, REMD and REST. The REMD ensemble data have already been published [212], while the REST method was run afterwards to compare the performance of the two enhanced sampling approaches.

The results for the ensemble sampling will be given separately in two sections.

7.3 REMD simulation

Using the REMD method, three simulation runs were performed with three different starting UII conformations referred to as *Omega I Open, Folded I* and *Lasso*. The *Omega I Open* and *Folded I* conformations were obtained from MD simulation from our collaborators [212]. A *Lasso* conformation was observed after initial runs of the first two REMD simulations were performed and analysed, revealing the appearance of another highly populated structure named as *Lasso*, which was then used to start another REMD simulation.

7.3.1 REMD simulation details

Three simulations were run for 500 ns each using the PMEMD module in AMBER 12 suite programs [125]. The temperature range was generated using the online temperature generator http://folding.bmc.uu.se/remd/ [178] with an overall expected acceptance ratio among replica of 30 % and provided us 64 replicas from 298 K to 543 K. The Amber ff99SB force field was used with explicit TIP3P water model [177]. The initial structures were solvated in a cubic box containing water molecules with periodic boundary conditions and neutralised with 1 Na^+ . The Particle Mesh Ewald [126] was used for the long-range interactions using a 10 Å cutoff. Bonds involving hydrogen were constrained using the SHAKE algorithm

[121] with a tolerance of 0.00001 Å. REMD simulations were performed in the NVT ensemble using a Langevin thermostat for the temperature coupling with a collision frequency of 1 ps^{-1}. 200 ps of NVT simulation was used to equilibrate the initial state to the desired temperature for each replica, following a rescaling of the velocities. Using these equilibrated replicas, 500 ns of REMD simulation was performed on each replica, consisting of 32 μs of molecular dynamics. All exchanges between neighbouring replicas were allowed every 2 ps in the NVT ensemble.

7.3.2 REMD simulation results

Three sets of the REMD simulations were performed to explore UII conformational ensemble. The initial 100 ns of each simulation were taken as equilibration time and were not included in the simulation analysis, as was done for the AVP and OT peptides.

Already established analysis procedure, consisting of analysing β-turn and γ-turn population, hydrogen bond population and cluster state diversity, was also performed for the UII peptide. The following sections provide more details.

7.3.2.1 β-turn and hydrogen bond population

Experimental data (Section 7.1) reported the UII conformational flexibility in terms of different β- and γ-turns. Here the β- and γ-turn population of the ring residues was explored using the definitions given in Section 1.2.1. The populations of different β-turns is given in the Table 7.1.

	6,7 type I	6,7 type II	7,8 type I	7,8 type II	8,9 type I	8,9 type II	8,9 type VII
Omega Open	22.23	2.28	8.46	0.68	8.91	1.78	1.64
Folded I	24.54	4.39	5.93	0.23	3.56	1.21	3.71
Lasso	31.6	2.31	8.95	0.13	4.43	0.44	4.57

Table 7.1: β-turn type populations from the three REMD simulations (Omega Open, Folded I and Lasso).

It shows that ring residues Phe^6, Trp^7, Lys^8 and Tyr^9 adopt a variety of β-turns. The most populated is the 6,7 β-turn followed by similarly populated 7,8 and 8,9 β-turns.

Next, the hydrogen bond population was also analysed between different combinations of residues that could make a hydrogen bond. The results summarised in Table 7.2 show that the most populated are $Cys^5O - Lys^8H$, $Phe^6O - Tyr^9H$ and $Trp^7O - Cys^{10}H$ hydrogen bonds. There is also a highly populated hydrogen bond between proline residue in N terminal tail and ring residue cysteine, $Pro^3O - Phe^6H$.

O - - H	**Folded I**	**Omega Open**	**Lasso**
$Thr^2 - Trp^7$	4.83	2.86	1.03
$Pro^3 - Phe^6$	32.79	21.55	25.76
$Cys^5 - Cys^{10}$	10.04	8.63	5.41
$Cys^5 - Lys^8$	19.25	18.75	21.00
$Phe^6 - Tyr^9$	9.08	11.67	13.05
$Phe^6 - Cys^{10}$	1.43	3.46	4.13
$Trp^7 - Tyr^9$	1.75	2.66	2.32
$Trp^7 - Cys^{10}$	5.43	10.08	6.13
$Tyr^9 - Asp^4$	1.39	2.27	1.61
$Tyr^9 - Cys^5$	1.58	2.16	1.54

Table 7.2: Different hydrogen bond populations from the three REMD simulations.

7.3.2.2 *cis/trans* Proline peptide bond

Since proline residue is present in the UII sequence at the third position in the N terminal tail, the *cis* population of the amide bond involving the nitrogen in the proline was also analysed. The *cis* bond was taken as adopting range of +/- 60 deg from the mean value of 0 deg. The results are given in Table 7.3, in which it can be seen that *cis* bond conformation is present in all simulation runs with a population between 1.5 - 3 %.

amide conformation	**Folded I**	**Lasso**	**Omega Open**
cis	2.94	2.07	1.49
trans	97.06	97.93	98.51

Table 7.3: The populations of the *cis/trans* amide bonds during the REMD simulations.

7.3.2.3 Torsion based clustering

Finally, the simulations were also analysed with the torsion based clustering software Dash. The ring torsion values $\psi 5$, $\phi\psi$ 6-9 and $\phi 10$ were extracted, and run through Dash software. Then, the sets of torsion values were compared between themselves with the *dashsim* program which calculates the circular similarity between the Dash states, and the final list of unique UII cluster states was obtained (Appendix C).

In total, the cluster states were divided into two group based on the adopted ring conformations, *open* or *folded*, giving in total 11 different UII substates. The substates were classified in terms of β-turn and hydrogen bond populations where the *open* substates were described as adopting 6,7 and 8,9 β-turns, and little or no hydrogen bonds were populated between their residues, while *folded* substates adopt 7,8 β-turns and are mostly stabilised by a number of hydrogen bonds (Figure 7.2).

The *open* state consists of the following substates, *Omega Open, Omega Open hbond, Omega II, Lasso, Scoop* and *Circle. Folded* ring cluster state include *Folded I, Folded II, Folded III, Folded IVb2* and *Inverted Folded* substates. The list of the dihedral angle specific for each state is given in the Appendix.

Each cluster substate is also defined by a unique set of ring torsion angles, therefore a circular similarity between the substates was performed using the *dashsim* program. Circular similarity analysis revealed that there is a clear distinction between different cluster substates since they are showing different circular similarity values (Table 7.4).

In the *open* states, the most similar substates are *Omega I Open* and *Omega I hbond* with a circular similarity value of 0.72, and *Omega I Open* and *Circle* with a circular similarity of 0.66 (Table 7.4 green cells).

From the *folded* states, the most similar substates are *Folded I* and *Folded II* (circular similarity = 0.75), and *Folded II* and *Folded III* (circular similarity = 0.68) (Table 7.4 pinks cells).

7.3.2.4 UII ensemble substate time distribution

The population of the identified individual cluster substates adopted by the UII peptide was plotted during the simulation time for all three simulation runs. Figure 7.3 shows that the most populated substate is *Lasso* (shown in yellow), which belongs to the *open* cluster state. In general, all cluster substates except *Lasso*, which is populated between 40 and 60 % during the simulation time, are populated

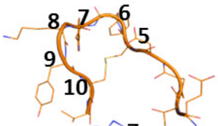
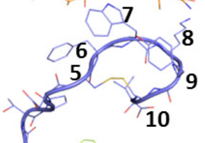
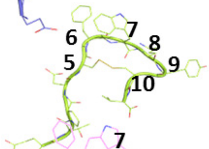
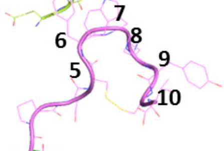
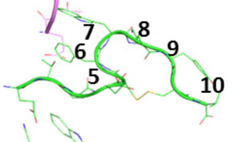
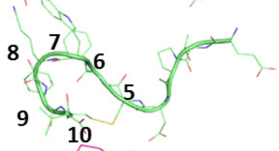
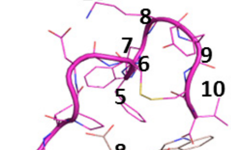
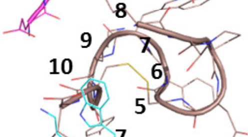
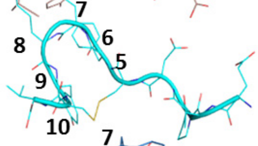
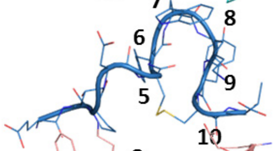
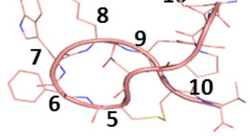
Cluster state	Cartoon representation	Hydrogen bond	Turn type
Omega I Open			8,9 β-I
Omega I hbond		^{7}O - ^{10}H	8,9 β-VIII
Omega II			8,9 β-II
Scoop		^{5}O - ^{8}H	6,7 β-I
Lasso			6,7 β-I
Folded I		^{6}O - ^{9}H	7,8 β-I
Folded II		^{5}O - ^{8}H,(^{9}H,^{10}H)	7,8
Folded III		^{5}O - ^{8}H(^{9}H,^{10}H)	7,8
Folded IVb2		^{6}O - ^{9}H	7,8 β-II
Inv Folded		^{6}O - ^{9}H ^{5}O - ^{8}H ,(^{10}H)	7,8
Circle			No turn

Figure 7.2: The UII cluster substates with associated hydrogen bond and turn type. If no hydrogen bond was characterised for a particular UII substate, then it is left blank space. A turn type is denoted with the turn centre residues.

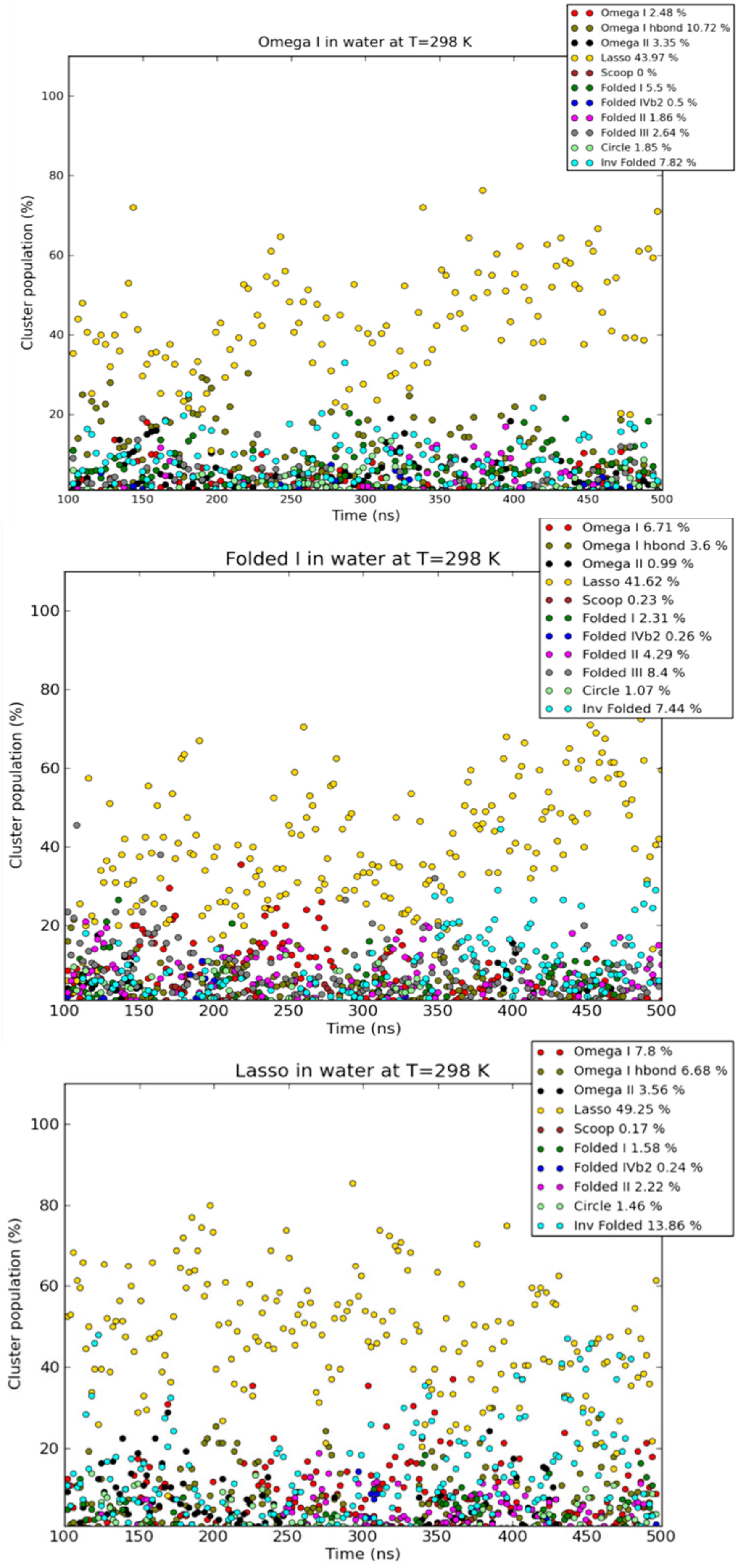

Omega I in water at T=298 K
Omega I 2.48 %
Omega I hbond 10.72 %
Omega II 3.35 %
Lasso 43.97 %
Scoop 0 %
Folded I 5.5 %
Folded IVb2 0.5 %
Folded II 1.86 %
Folded III 2.64 %
Circle 1.85 %
Inv Folded 7.82 %
Cluster population (%)
100
80
60
40
20
100
150
200
250
300
350
400
450
500
Time (ns)
Folded I in water at T=298 K
Omega I 6.71 %
Omega I hbond 3.6 %
Omega II 0.99 %
Lasso 41.62 %
Scoop 0.23 %
Folded I 2.31 %
Folded IVb2 0.26 %
Folded II 4.29 %
Folded III 8.4 %
Circle 1.07 %
Inv Folded 7.44 %
Cluster population (%)
100
80
60
40
20
100
150
200
250
300
350
400
450
500
Time (ns)
Lasso in water at T=298 K
Omega I 7.8 %
Omega I hbond 6.68 %
Omega II 3.56 %
Lasso 49.25 %
Scoop 0.17 %
Folded I 1.58 %
Folded IVb2 0.24 %
Folded II 2.22 %
Circle 1.46 %
Inv Folded 13.86 %
Cluster population (%)
100
80
60
40
20
100
150
200
250
300
350
400
450
500
Time (ns)

Figure 7.3

	Omega I Open	Omega I hbond	Omega II	Folded I	Folded II	Folded III	Folded IVb2	Inv Folded	Lasso	Scoop	Circle
Omega I Open	1.00	0.72	0.51	0.50	0.41	0.45	0.50	0.38	0.51	0.60	0.66
Omega I hbond	0.72	1.00	0.56	0.58	0.50	0.46	0.62	0.42	0.46	0.53	0.54
Omega II	0.51	0.72	1.00	0.36	0.33	0.33	0.43	0.49	0.55	0.38	0.41
Folded I	0.50	0.58	0.36	1.00	0.75	0.56	0.61	0.43	0.45	0.49	0.53
Folded II	0.41	0.50	0.33	0.75	1.00	0.68	0.54	0.41	0.38	0.40	0.45
Folded III	0.45	0.46	0.33	0.56	0.68	1.00	0.44	0.47	0.42	0.49	0.48
Folded IVb2	0.50	0.62	0.43	0.61	0.54	0.44	1.00	0.32	0.34	0.52	0.31
Inv Folded	0.38	0.42	0.49	0.43	0.41	0.47	0.32	1.00	0.47	0.41	0.46
Lasso	0.51	0.46	0.55	0.45	0.38	0.42	0.34	0.47	1.00	0.54	0.65
Scoop	0.60	0.53	0.38	0.49	0.40	0.49	0.52	0.41	0.54	1.00	0.65
Circle	0.66	0.54	0.41	0.53	0.45	0.48	0.31	0.46	0.65	0.65	1.00

Table 7.4: The circular similarity between different UII cluster substates.

not more than 20 % over all simulation repeats. However, in REMD Omega I and Folded I simulations there is some indication that the population of *Lasso* is gently increasing which suggests that it is perhaps not quite converged.

Beside *Lasso*, other highly populated *open* substates are *Omega I Open* and *Omega I hbond* taken together because of the high circular similarity between their ring residues (Table 7.4). From the *folded* state, the most populated substate is *Inverted Folded*, between 5 - 10 % .

The UII substate populations obtained with REMD cannot be compared with MD populations because they could not report the populations of the substates due to the very slow UII conformational dynamics on the MD timescale (see Section 7.1) [212].

7.3.2.5 The population of *cis* amide in UII cluster states

Indicated in Table 7.3 $cis - Pro^3$ amide appeared in all simulation ensembles. Then it was examined if it is selective for any of the UII cluster substates. Table 7.5 shows that all UII substates, except *Scoop*, are adopting the $cis - Pro^3$ isomer state. It is the most populated in the *Lasso* substate, followed by *Omega I Open, Omega I hbond, Inverted Folded.* At the same time, these are also the most populated substates in the REMD simulations, suggesting that there is no particular preference for the *cis* amide to be associated with a particular ring conformation.

	Folded I	**Lasso**	**Omega I Open**
Omega I Open	6.75	14.46	5.15
Omega I hbond	7.67	5.02	8.41
Omega II	1.61	1.71	1.54
Folded I	0.75	0.73	3.44
Folded II	2.02	0.73	1.72
Folded III	3.40	0.0	1.03
Folded IVb2	0.0	0.24	0.86
Inv Folded	10.14	4.78	8.41
Lasso	48.27	51.71	53.95
Scoop	0.0	0.0	0.0
Circle	0.0	1.83	0.68

Table 7.5: The $cis - Pro^3$ amide bond population in the different UII cluster substates in the three REMD simulation repeats.

7.4 REST simulation

In order to explore the efficiency of the REMD method, UII peptide was also run with the REST method. As shown for the example of AVP and OT peptides (Chapters 4, 5), the REST method produced converged conformational sampling with far fewer number of replicas, reducing the computational cost and real time needed to get converged sampling. Here the practical and conformational sampling advantage of REST method over REMD method will be explored as well, but on the same peptide.

7.4.1 REST simulation details

The same three UII cluster substates used to perform REMD simulations were also used as initial conformations to run REST simulations. The method was run in Gromacs with the Plumed patch [155]. The Amber ff99SB was used with explicit TIP3P water model [177]. The system was neutralised with the Na^+ counterion. The Particle Mesh Ewald [126] was used for the long-range interactions using a 10 Å cutoff. Bonds involving hydrogen were constrained using the SHAKE algorithm [121] with a tolerance of 0.00001 Å. REST simulations were performed in the NVT ensemble using a Langevin thermostat for the temperature coupling with a collision frequency of 1 ps^{-1}.

The simulations were run for 300 ns using 12 replicas in the temperature range 298 K - 900 K. The replicas were geometrically distributed to give the acceptance ratio between 20 and 35 %.

7.4.2 REST simulation results

7.4.2.1 β-turn and hydrogen bond population

The β-turn population analysis was done in the same way as for the REMD simulations (Section7.3.2.1). The populations of different β-turns are given in the Table 7.6 showing that UII peptide prefers ring conformations with β-turns centred at residues Phe^6, Trp^7, Lys^8 and Tyr^9.

Next, the hydrogen bond population was also analysed between different combinations of residues that could make a hydrogen bond. The results summarised in Table 7.7 show that the most populated intracyclic hydrogen bonds are $Cys^5O -$

	6,7 type I	6,7 type II	7,8 type I	7,8 type II	8,9 type I	8,9 type II	8,9 type VII
Omega Open	24.89	2.23	6.57	0.45	9.73	0.46	2.24
Folded I	22.93	2.17	10.58	0.25	8.45	1.41	1.96
Lasso	25.28	0.75	6.61	0.63	6.94	1.60	1.03

Table 7.6: β-turn type populations from the three REST simulations

Lys^8H, $Phe^6O - Tyr^9H$ and $Trp^7O - Cys^{10}H$, while also highly populated is the bond between tail and ring residues $Pro^3O - Phe^6H$.

O − H	Folded I	Omega Open	Lasso
$Thr^2 - Trp^7$	3.02	3.54	2.00
$Pro^3 - Phe^6$	20.45	17.57	18.24
$Cys^5 - Cys^{10}$	3.55	6.91	5.19
$Cys^5 - Lys^8$	15.45	17.88	15.78
$Phe^6 - Tyr^9$	8.72	10.13	9.05
$Phe^6 - Cys^{10}$	0.12	1.27	2.14
$Trp^7 - Tyr^9$	2.02	1.85	1.42
$Trp^7 - Cys^{10}$	8.63	10.13	6.77
$Tyr^9 - Asp^4$	1.12	1.01	2.74
$Tyr^9 - Cys^5$	1.44	1.97	2.04

Table 7.7: Different hydrogen bond populations from the three REST simulations.

7.4.2.2 *cis/trans* Proline peptide bond

Table 7.3 shows that the $cis - Pro^3$ amide bond was populated between 1.5 to 3 % in the REMD simulations. The *cis* amide population during the REST simulations was also examined. A Table 7.8 is showing that in the REST simulation a higher percentage of *cis* amide bond is observed than in REMD simulations, ranging between 4 and 7 %. A maximum *cis*-Pro population of 10% was suggested by experimental data [212].

amide conformation	Folded I	Lasso	Folded I
cis	3.91	4.74	6.46
trans	96.09	95.26	93.54

Table 7.8: The populations of the *cis/trans* amide bonds during the REST simulations.

7.4.2.3 Torsion based clustering

The population of the identified UII substates was examined in the REST simulations too. Analysing the same ring torsion angles with the Dash software as was done with REMD simulations, a final list of UII cluster states was obtained, then plotted against simulation time. Figure 7.4 shows that the most populated conformational state is the *Lasso* substate followed by the *Omega I Open* substates. A *folded* substates *Folded I* and *Inverted Folded* were also highly populated, up to 10 %, while other *open* and *folded* substates were less well populated.

7.4.2.4 The population of *cis* amide in UII cluster states

As given in the Table 7.8, $cis - Pro^3$ amide bond appeared between 4 % to 7 % in the REST simulation trajectories. A further analysis on the individual cluster substates revealed that the *cis* amide is not selective for any UII cluster substate (Table 7.9). It also showed that it appeared with highest population in *Lasso* substate followed by *Omega I Open, Omega I hbond, Omega II, Folded I and Inverted Folded* substates. The similar result was obtained with REMD simualtions (Table 7.5), although in REMD simulations the $cis - Pro^3$ population was lower in *Omega II* and *Folded I* substates.

	Folded I	**Lasso**	**Omega Open**
Omega I Open	2.27	7.65	3.94
Omega I hbond	14.57	6.92	7.25
Omega II	13.43	8.81	0.14
Folded I	6.21	4.08	7.71
Folded II	1.14	0.0	0.51
Folded III	1.13	1.94	3.28
Folded IVb2	0.0	0.52	0.67
Inv Folded	3.01	3.46	5.00
Lasso	52.21	49.68	73.86
Scoop	0.0	0.00	0.0
Circle	0.13	3.14	0.0

Table 7.9: The $cis - Pro^3$ amide bond population in the different UII cluster substates in the three REST simulation ensembles.

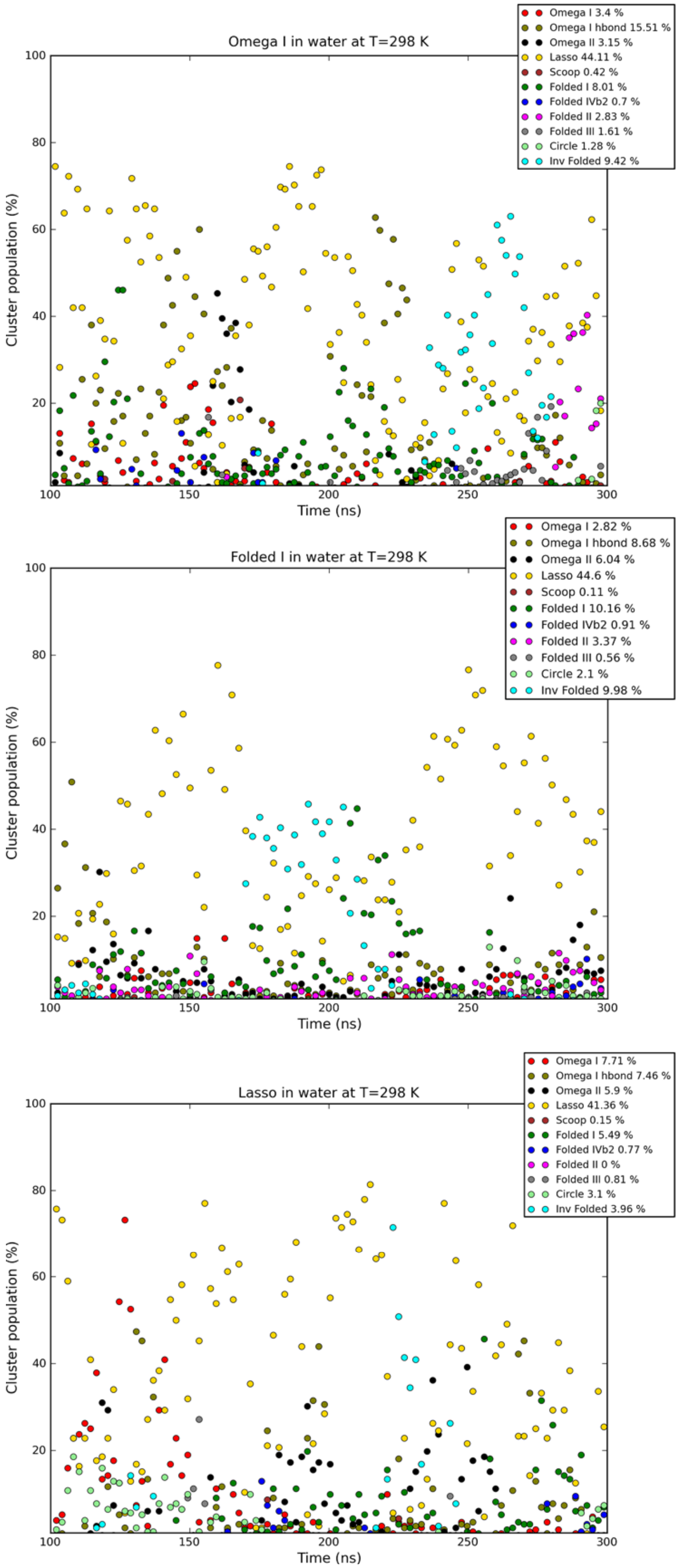


Figure 7.4: The time distribution of individual UII cluster substates in the three REST simulations.

7.5 REMD vs. REST conformational ensemble

Finally, the REMD and REST simulation performance in terms of conformational sampling will be compared in this Section.

The relative populations of Urotensin II peptide conformational ensemble substates from REMD and REST simulation repeats are compared in Figure 7.5.

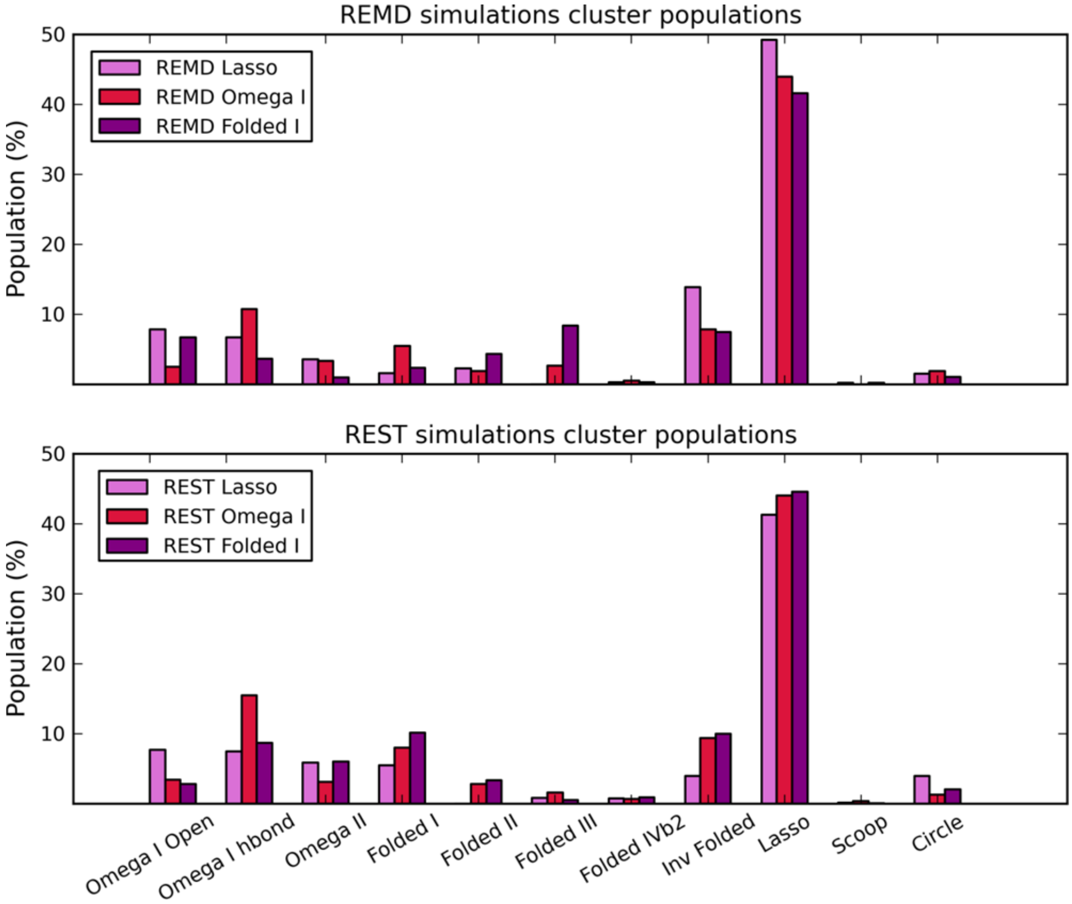


Figure 7.5: The cluster substate populations from the three simulation repeats from *a*) REMD and *b*) REST simulations.

Both methods predict that *Lasso* is the most populated substate with average population between 40 % and 50 % in all performed simulations. The second most populated substates are *Omega I Open* and *hbond*, together with *Inverted Folded*. The circular similarity of 0.72 between *Omega I Open* and *Omega I hbond* suggests that these two substates can be considered as one substate.

Although the individual populations of these three most populated substates differ between the simulation repeats and methods, their overall population agree well. *Omega I Open* is populated 5 - 10 %, *Omega I hbond* 8 - 15 % and *Inverted Folded* is between 5 % and 10 %.

Next, in the REST simulations, there is a higher percentage of *Omega II* and *Folded I* substates, compared to REMD simulations, while *Folded IVb2, Scoop* and *Circle* are the lowest populated substates in all simulation repeats.

The strongest disagreement between the methods and intra-method simulation repeats is shown for the *Folded III* substate. It is highly populated in the REMD Folded I simulation only, while in the other simulation repeats, it was not populated more than 3 %.

Overall, enhanced sampling REMD and REST simulations show that Urotensin II peptide is a flexible peptide, most of the simulation time preferring the *open Lasso* substate. The 6,7 β-turn population, characteristic for *Lasso* substate, is populated 25 % - 35% over all simulation repeats, agreeing well with the *Lasso* total population 40 % - 50 %. This substate was not characterised by any hydrogen bond. The NMR data reported only on the widened 7,8,9 γ-turn.

Other highly populated *open* substates are *Omega I Open* and *Omega I hbond.* These substates are 72 % similar in terms of ring torsion angles (Table 7.4), so their populations can be looked at together (Figure 7.5). They are populated approximately between 15 - 25 %, and characterised with 8,9 β-turns types I and VII. In both methods, these β-turns were populated 8 - 15 %. The hydrogen bond specific for this substates $Trp^7 - Cys^{10}$ was adopted 5 - 10 %, depending on the simulation (Tables 7.2 and 7.7).

The *Omega II open* substate was populated up to 5 % in all simulation repeats (Figure 7.5) with characteristic β-turn type VII similarly populated up to 3 % (Tables 7.1 and 7.6). This substate was not characterised by any hydrogen bond.

The two most populated *folded* substates were *Folded I* and *Inverted Folded.* A *Folded I* substate, which mostly adopted 7,8 β-turn type I, was populated more in REST simulations (approximately 10 %) than in REMD simulations (approximately 5 %). Another highly populated *folded* substate *Inverted Folded* was similarly populated in simulations performed by both methods (up to 10 %). In both substates the $Phe^6O - Tyr^9H$ hydrogen bond, populated between 8 % - 13 %, appeared (Tables 7.2 and 7.7).

All *folded* substates were described as stabilised by different hydrogen bonds (Figure 7.2), therefore the population of $Cys^5O - Lys^8H$ ranging between 15 % -

22 % is agreeing well with overall folded substate populations of approximately 20 %. Other intra-cyclic hydrogen bonds not emphasised in this Section are minorly populated, not more than 3 % (Tables 7.2 and 7.7).

Summary. The comparison between the torsion based UII cluster substates and the population of hydrogen bonds and β-turns during the analysed simulation time revealed that their populations are similar enough to consider our simulation repeats as converged, for both methods. The most populated UII substate known as *Lasso* was almost three times more populated than any other individual UII substate. The total population of all substates in the simulation was approximately 80 %, the remaining 20 % were considered as transient substates which could not be assigned to any of the substate representatives, and showed no similarity between themselves.

7.5.1 The population of $cis - Pro^3$ amide bond

Regarding the populations of the $cis - Pro^3$ amide bond, it was more populated in the REST (4-7 %) than in the REMD simulation ensembles (1.5-3 %). This suggests that the energy barrier of *cis/trans* transition is more easily overcomed in the REST approach of scaling the certain interactions in the system than with REMD method, where maybe the temperature of 550 K was not high enough to see more frequent *cis/trans* interconversion, or perhaps a longer simulation time is required. The reported experimental population was around 10 % [212].

7.5.2 Comparison with experimental data

The substate description can also be compared with the known experimental data (see Section 7.1). The NMR description of UII conformations in aqueous solution as adopting turns centred at residues Lys^8, Tyr^9 resembles our *open* ring substates, while the NMR description of UII ring conformation in SDS micelles as *folded* with turn at residues 7, 8 agrees with our description of *folded* substates which were also described as all adopting turns centred at residues Trp^7 and Lys^8.

7.5.3 The N-terminal tail

The N-terminal tail conformation was described as flexible by the experimental data. A tail residue Pro^3 showed a high preference to make hydrogen bond with the carbonyl oxygen of the ring residue Phe^6. The same population of this hydrogen bond in the range of 20 % - 30 % across all simulation ensembles was observed, suggesting that the N terminal tail was pointing towards the ring almost one third of the simulation time. There also shortly appeared another hydrogen bond between $Thr^2 - Trp^7$ for up to 5 %. These data suggests that a four residue N terminal tail is flexible enough to make hydrogen bonds with the ring part of the structure further stabilising UII conformation.

7.6 UII chemical shifts

After the extensive analysis of the UII conformational ensemble, the data were validated against the proton chemical shifts obtained at pH 6.0 and temperature 298 K [212].

The proton chemical shifts were calculated using Gaussian09 software [116] with B3LYP/6-31G(d) level of DFT theory (see Section 3.2). The procedure applied for UII cluster representatives is the same as already described in Sections 4.7.0.2 for AVP and then applied to all the peptides studied in this work.

7.6.0.1 The choice of the representative structures

The representative structures were chosen to fulfil the same conditions as for AVP and OT; to be scattered in approximately equal intervals along the trajectory, and to be within the 1 SD of the torsion angle distribution mean (Appendix C.1, C.2). There are five representative structures for each UII representative structure substate, but *Scoop* for which four structures were extracted from the REMD trajectory.

First, the variance within calculated chemical shifts for all representative structures within each cluster substates was checked. Figure 7.6 shows that all cluster substates adopt a tight range of chemical shift values, with an exception of Lys^8 HA for Circle substate and HE2 for Inverted Folded.

Next, the statistical analysis was used to check the peptide intra-substate chemical shift variance. Three statistical measures of similarity MUE, MSE and R^2 (Section 3.5.3) were used, and were bootstrapped afterwards to account for the fact that only a few (4 or 5) structures were chosen as a representative of the several hundred frames belonging to a particular cluster.

The intra-cluster R^2 variance was plotted for the *open* and *folded* cluster states shown in Figure 7.7. When the R^2 values are overlayed, no significant difference between the structures belonging to a particular cluster is observed. The *open* state structures (upper part of Figure 7.7) show wider R^2 ranges, mostly due to the *Circle* conformation which has lower R^2 compared to other *open* states. While for most states R^2 falls within the range 0.94 to 0.97, the best agreement with experimental data are for *Lasso* and *Omega I Open* substates of the *open* cluster group, and *Folded I* and *Folded IVb2* from *folded* cluster state.

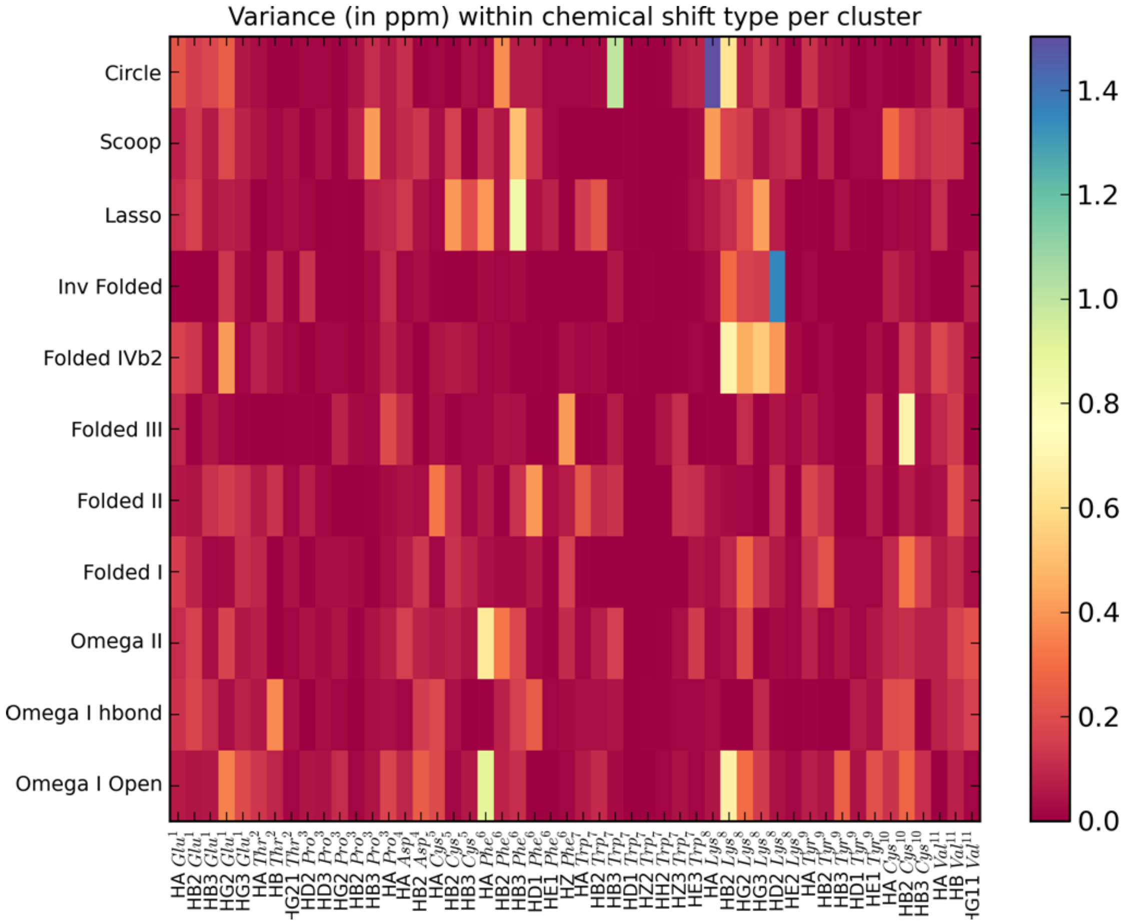


Figure 7.6: The variance within chemical shift types for all cluster substate representative structures.

7.6.1 Bootstrapping of the individual cluster states

Since each cluster state was represented by 5 (4 for *Scoop*) structures, and some states, especially the highest populated ones, have several hundreds of frames belonging to them, to ensure that the picked frames are truly representative of the cluster state, the shifts from structures within each cluster were bootstrapped in such a way that values of the individual shifts from each structure for each proton shift were selected one at a time to build 10k shift sets, and then each of these was compared with the experimental values. The results of this analysis are shown in Figure 7.8 and summarised in Table 7.10.

The best agreement with experimental data is shown for substates belonging to both *open* and *folded* cluster states. A *Folded I* substate is the best performing, after which follow *Omega I Open, Folded IVb2* and *Lasso*, all three with overlap-

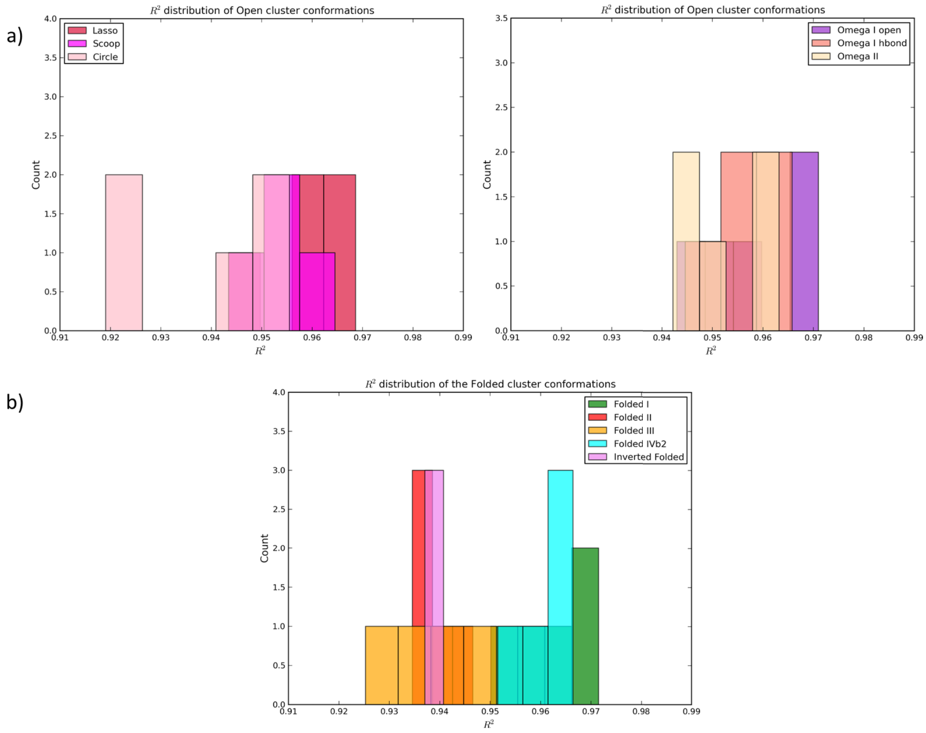


Figure 7.7: The R^2 distribution of the representative structures for the each of the UII cluster state. Upper part *a*) shows the distribution of the *Open* cluster states, and *b*) shows the distribution of the *Folded* cluster state.

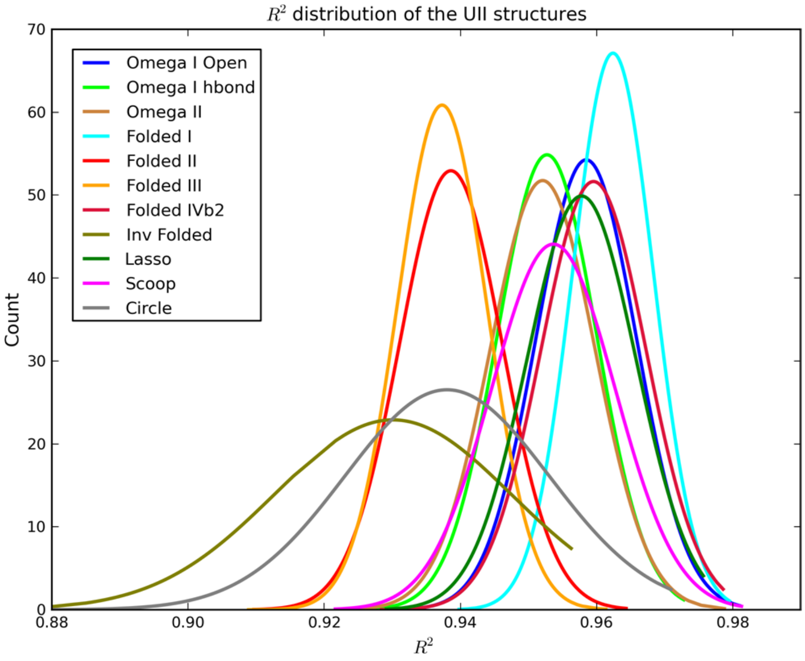


Figure 7.8: The bootstrapped R^2 distribution of the all UII cluster substates.

ping R^2 distributions (Figure 7.8).

On the other side, the weakest agreement with experimental data show mostly *folded* substates *Folded II, Folded III, Inverted Folded* and an *open* structure, *Circle.*

Figure 7.8 also shows that a few substates are showing overlapping R^2 distribution, here given in the order from the highest substate R^2 values to the smallest:

- Folded I
- Omega I Open, Folded IVb2, Lasso
- Omega I hbond, Omega II, Scoop
- Folded II, Folded III, Circle
- Inverted Folded

The exact ranges of the bootstrapped metrics are given in the Table 7.10, where it can be seen that for the best performing states, *Omega I Open* has the lowest MUE value of all states, followed by *Lasso, Folded I* and *Folded IVb2.*

	MUE	**MSE**	R^2
Omega I Open	0.154 < 0.189 < 0.238	-0.107 < -0.041 < 0.022	0.93 < 0.96 < 0.98
Omega I hbond	0.219 < 0.279 < 0.355	-0.154 <-0.050 < 0.049	0.93 < 0.95 < 0.97
Omega II	0.211 < 0.260 < 0.321	0.083 < -0.094 < -0.010	0.93 < 0.95 < 0.97
Folded I	0.197 < 0.230 < 0.287	-0.106 < -0.029 < 0.049	0.94 < 0.97 < 0.98
Folded II	0.250 < 0.311 < 0.394	-0.097 < 0.012 < 0.125	0.91 < 0.93 < 0.96
Folded III	0.269 < 0.341 < 0.444	-0.091 < 0.026 < 0.164	0.91 < 0.93 < 0.96
Folded IVb2	0.194 < 0.230 < 0.301	-0.029 < 0.053 < 0.140	0.93 < 0.96 < 0.98
Inv Folded	0.299 < 0.366 < 0.461	-0.134 < -0.008 < 0.126	0.89 < 0.93 < 0.96
Lasso	0.175 < 0.220 < 0.279	-0.082 < -0.003 < 0.080	0.93 < 0.96 < 0.98
Scoop	0.173 < 0.232 < 0.294	-0.132 < -0.041 < 0.040	0.92 < 0.95 < 0.97
Circle	0.213 < 0.272 < 0.350	-0.098 < -0.002 < 0.105	0.89 < 0.94 < 0.97

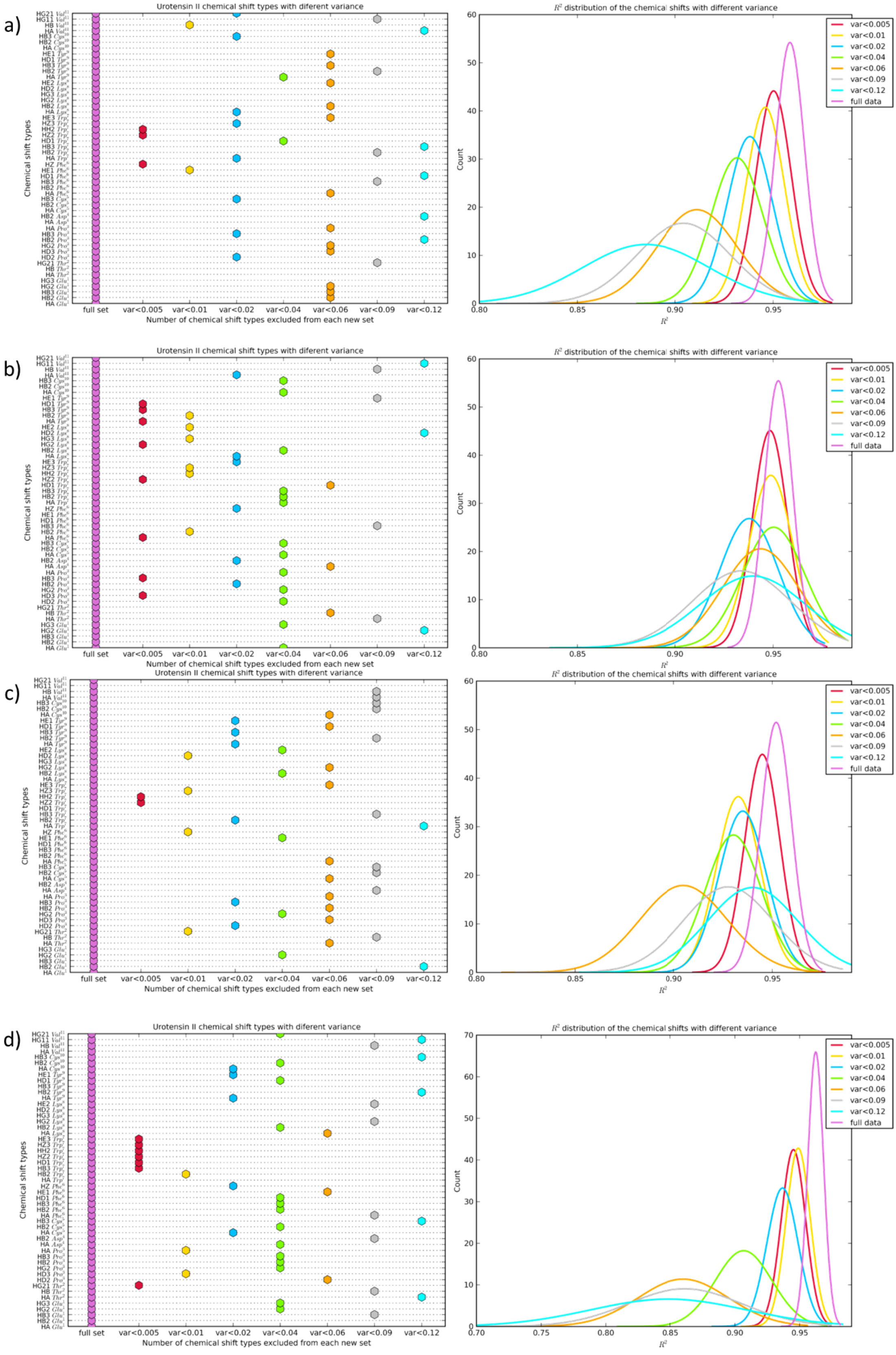
Table 7.10: The bootstrapped values of the three statistical measures of similarity, Mean Unsigned Error (MUE), Mean Signed Error (MSE) and coefficient of determination (R^2) for different UII cluster members

7.6.1.1 Analysis of the individual chemical shift types

The analysis of the individual chemical shift types was performed to monitor the weight of the particular chemical shift types to the overall R^2 distribution. It gives the idea of the chemical shift types which upweigh or downweight the final distribution.

The chemical shift types with variance lower than 0.005 ppm, 0.01 ppm, 0.02 ppm, 0.04 ppm, 0.06 ppm, 0.09 ppm and 0.12 ppm were extracted in the subsequent way, and the R^2 distribution was plotted with the remaining number of chemical shift types.

Figure 7.9 a) - b) shows that Omega Open, together with Omega II cluster substate (7.9 c)) shows very good agreement with experimental data, with R^2 distribution not lower than 88 % for almost all chemical shift type combinations, which contains analysis on the almost 2/3 of all chemical shift types. The same pattern was also observed for *Lasso* cluster (Figure 7.9 d)), while *Folded I* cluster substate showed wider variance (Figure 7.9 e)), where chemical shift types with variance higher than 0.12 ppm exhibited the lowest correlation with experimental data, which suggests that these chemical shift types down-weight R^2 distribution.


a)
b)
c)
d)
Urotensin II chemical shift types with diferent variance
Chemical shift types
full set
var<0.005
var<0.01
var<0.02
var<0.04
var<0.06
var<0.09
var<0.12
Number of chemical shift types excluded from each new set
R^2 distribution of the chemical shifts with different variance
Count
R^2
var<0.005
var<0.01
var<0.02
var<0.04
var<0.06
var<0.09
var<0.12
full data
0.70
0.75
0.80
0.85
0.90
0.95
0
10
20
30
40
50
60
70

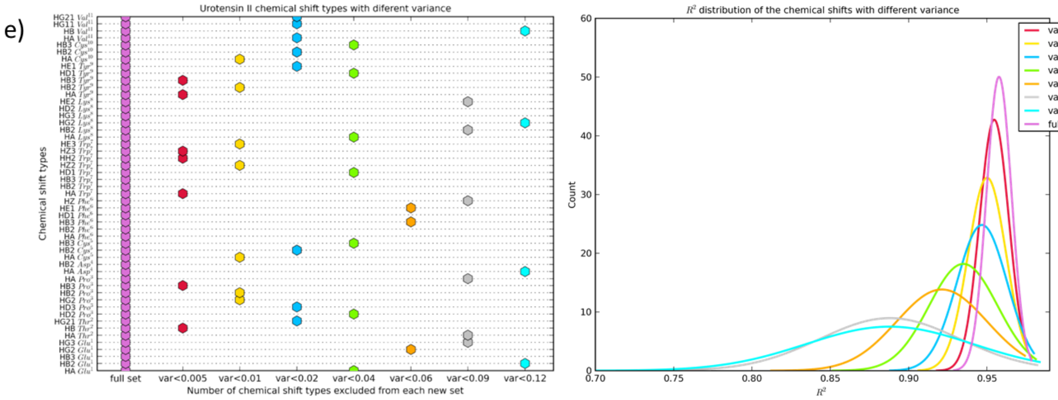


Figure 7.9: The R^2 distribution of the theoretically calculated vs. experimental chemical shifts with different chemical shift data sets (right plots) depending on the value of the chemical shift type variance excluded from final analysis (on the left plots). Distributions are plotted separately for a) Open Omega, b) Open Omega hbond, c) Omega II, d) Lasso, e) Folded I Urotensin II cluster substate.

7.6.2 Ensemble model

The ensemble model was built by weighting each chemical shift with the normalised population of the particular cluster substate, and then summing over all substates. The model was previously introduced in the Chapter 4.7.2, and here only the results will be given.

The conformational ensemble of the UII peptide was determined using two enhanced sampling methods, REMD and REST, each run using three different starting conformations (*Omega I Open, Lasso* and *Folded I*). The comparison between the cluster member populations in more detail is explained in Section 7.5. Thus here only the equilibrium model equation will be presented for one REMD simulation repeat, but was applied in the same way to all method repeats.

For the example of the REMD Folded I simulation, the ensemble equilibrium model equation 7.6.2 is showing that each chemical shift value was multiplied with the population of the cluster substate, where numbers 1, 2... 11 are following this order of the substate populations, *Omega I Open, Omega I hbond, Omega II, Folded I, Folded II, Folded III, Folded IVb2, Inverted Folded, Lasso, Scoop* and *Circle*. The time evolution of the substate populations can be examined in Figure 7.3.

$$\delta_{eq} = 0.088 * \delta_1 + 0.046 * \delta_2 + 0.013 * \delta_3 + 0.032 * \delta_4 + 0.055 * \delta_5 + 0.109 * \delta_6 \\ +0.003 * \delta_7 + 0.096 * \delta_8 + 0.542 * \delta_9 + 0.003 * \delta_{10} + 0.013 * \delta_{11}$$

Equation 7.6.2 states that the value of the particular ensemble chemical shift is obtained as a weighted sum of the individual shifts of each of the conformers $\delta_{1,2...11}$. It assumes fast dynamics on the NMR time scale. Since each state consists of several structures, the model was built in such a way that one structure at the time belonging to a particular substate was extracted at random, and then multiplied by the population of that substate. The result for the bootstrapped R^2 distribution is shown in Figure 7.10.

The R^2 REMD and REST ensemble histograms given on the Figure 7.10 show that ensemble distributions have almost identical overlap with R^2 values in the range 0.975 to 0.985, with the exception of REMD Lasso simulation which shows a bit lower R^2 distribution range until 0.970.

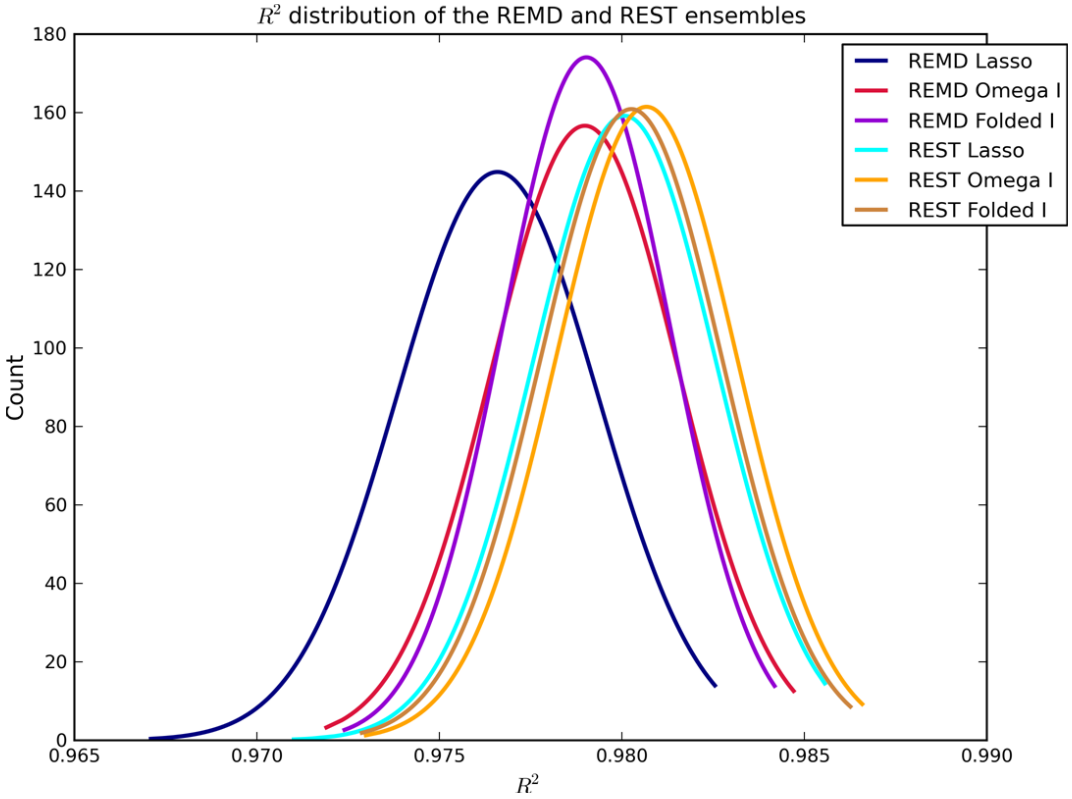


Figure 7.10: The histogram of the bootstrapped R^2 ensemble values for REMD and REST simulation runs.

If ensemble model is compared with the values of R^2 obtained from the bootstrapping analysis of the individual UII substates, then it can be observed that ensemble model (Figure 7.10) is adopting higher R^2 values than individual substates (Figure 7.8).

Next, if we take a look at the population of the individual cluster members (Figure 7.5), some states which are low populated in the simulations have very good agreement with experimental data, such as *Folded IVb2* and *Scoop*. This would imply that the contribution of these states to the model would be down weighted at the end, while the contribution of the states which are more populated but agree less with the experimental data, will be higher. Overall, *open* states *Omega I Open, Omega I hbond, Omega II* together with *folded* states *Folded I* and *Lasso* will contribute the most to the final model R^2 distribution because they are highly populated in all simulations, and adopt higher R^2 values than other substates (Figure 7.8).

Finally, the individual ensemble R^2 distributions will be discussed in more detail. From Figure 7.10 can be observed that REST simulations named as Omega I, Folded I and Lasso have overlapping distributions with the bootstrapped R^2 means centred at 0.981, which suggests very good convergence of the simulation ensembles. The ratio of the individual substates with the highest individual R^2 (Figure 7.8) in these ensembles are given below. The best agreement ensemble is given in bold.

- **REST Omega I**: Omega Open - 19 % : Lasso - 44 % : Folded I - 8 %
- REST Folded I: Omega Open - 10 % : Lasso - 44 % : Folded I - 10 %
- REST Lasso: Omega Open - 14 % : Lasso - 44 % : Folded I - 8 %

Just as the REST ensembles are adopting almost the same bootstrapped mean R^2 values (Figure 7.10), the ratio between the most populated substates is also very similar, as would be expected. However, in the best performing REST ensemble in terms of the highest R^2 value (REST Omega I), Omega Open substates (*Omega I Open* and *Omega I hbond*) are a bit higher populated compared to other ensembles.

Next, two REMD simulations named as Folded I and Omega I also adopt overlapping distributions, but with bootstrapped R^2 mean at 0.978. When compared, the main substates ratio for these simulations are

- REMD Folded I: Omega Open - 10 % : Lasso - 42 % : Folded I - 3 %

- REMD Omega I: Omega Open - 13 % : Lasso - 44 % : Folded I - 6 %

Overall comparison between all ensembles suggests that the main contribution to the ensemble bootstrapped R^2 value is the difference between populations of *Omega Open* and *Folded I* substates. In the REST ensembles which all have almost identical R^2 peak at 0.981, there is a slightly higher population of the *Folded I* substate compared to the REMD simulation ensembles.

7.6.2.1 Optimal population ratios

This analysis was performed to check the population ratios which would give the best agreement with the experimental chemical shift data ($R^2 > 0.99$). The obtained population ratios (Figure 7.11) suggest that the simulation data give very good approximation of the cluster populations for the given set of experimental chemical shifts. The most observed cluster substates belong to *Omega Open, Lasso* and *Folded I* which matches the simulation data although *Lasso* was populated approximately 40 % in all simulation repeats, while in the optimal population ratio calculations it is populated 5 - 20 %.

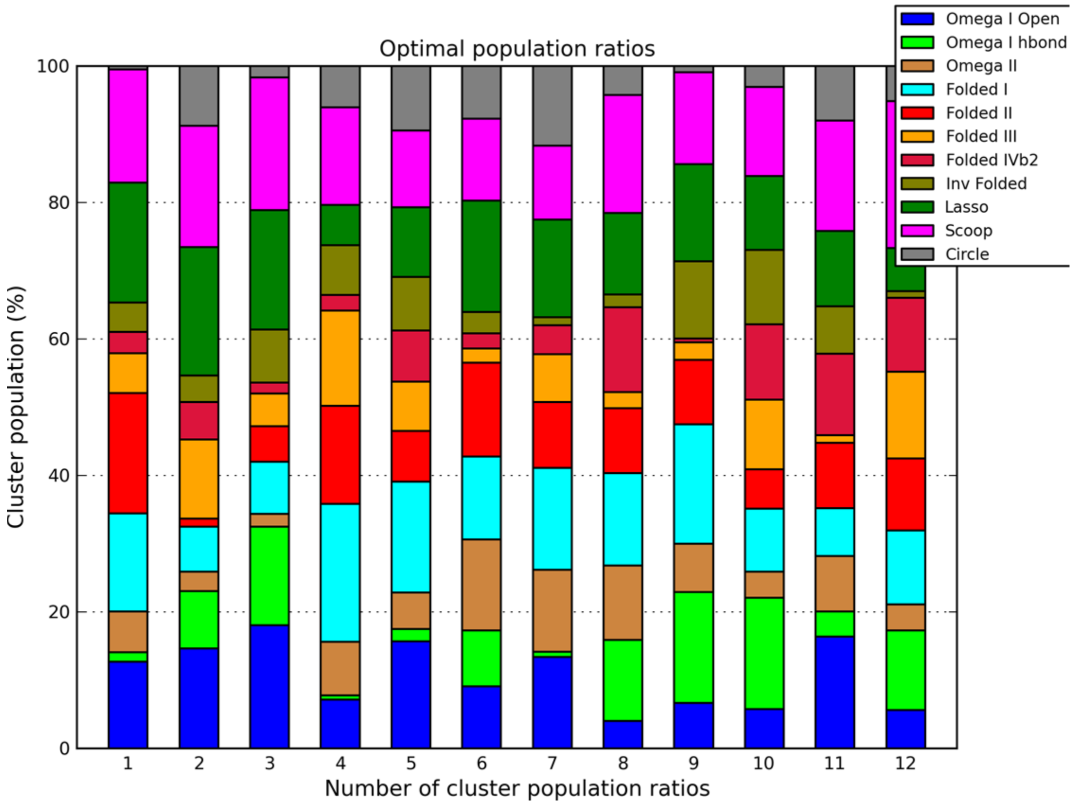


Figure 7.11: The cluster substate population ratios which have $R^2 > 0.99$ when compared against experimental chemical shift data.

7.7 Conclusions

Two enhanced sampling methods, REMD and REST, were applied to study the conformational ensemble of Urotensin II peptide. First, the performance of the methods was compared, and then the obtained conformational ensembles were validated against experimental data.

The conformational ensemble was extensively studied by running in total six simulations, three repeats for each enhanced sampling method using the same starting conformations. Our results suggest that converged conformational sampling were obtained with both methods, but with a significant computational cost using temperature Replica Exchange. In the Solute Tempering method, five times fewer number of replicas was used without affecting the sampling efficiency. Furthermore, the UII substate ratios between methods and simulation repeats suggests that a rather complete picture of the UII conformational ensemble was obtained. The experimental data, although rare, were well reproduced in terms of secondary structure motifs. They reported on the structures with turns centred at residues 8,9 which resemble our *open* ring state types, while *folded* conformation observed in SDS micelles resemble our *folded* ring substates.

Next, the ensemble validation against experimental chemical shift data revealed that the equilibrium ensemble model, already tested on AVP and OT peptides, was also performing better than the individual ensemble substates in the case of the UII peptide too. The simulation ensembles, together with the bootstrapped R^2 analysis revealed that UII is a flexible peptide adopting two major ring conformations, *open* and *folded*.

Chapter 8

Urotensin Related Peptide

Urotensin Related Peptide (URP) is a hormone peptide analogue of Urotensin II peptide [213]. They share the same structural motif of a six membered ring closed by disulphide bridge between two Cystein residue ($Cys^2 - Cys^7$). URP differs from UII only in the shorter N terminal, made of single alanine residue which contributes to the total charge of +1 at pH 4 - 8.

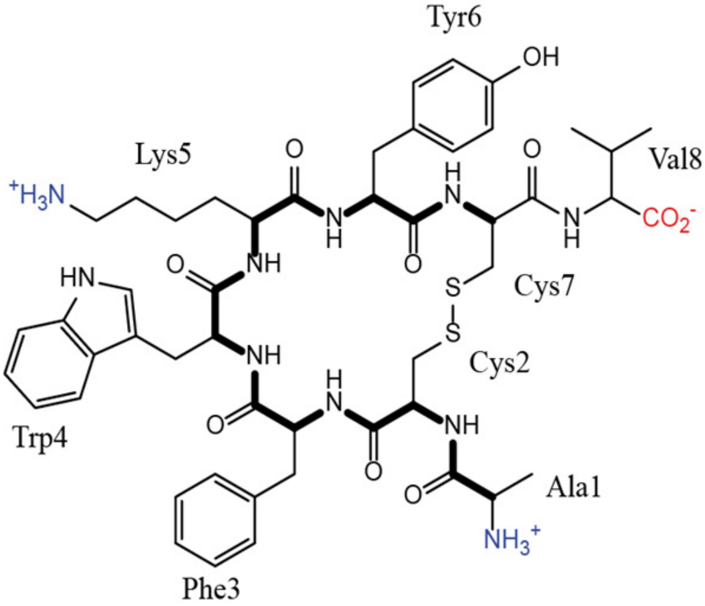


Figure 8.1: The structure of the Urotensin Related Peptide. The URP sequence contains eight residues, $Ala^1 - Cys^2 - Phe^3 - Trp^4 - Lys^5 - Tyr^6 - Cys^7 - Val^8$.

8.1 Known structural data

Just like AVP and OT are often studied together, the same is true of Urotensin II and Urotensin Related Peptide. **The experimental data** relating to URP structure are similarly as rare as for UII. The structure description varies from unstructured ring conformation [51] to turns centred at residues Lys^5 and Tyr^6 described by the 4,5,6 γ'-turn and $Trp^4O - Tyr^6H$ hydrogen bond [214].

The study performed by Brancaccio et al. [51] suggested high structural similarity between UII and URP ring conformations.

The NMR study in SDS micelles of the URP-like $UII_{(4-11)}$ peptide reported a 7,8 (4,5) β-turn type II' conformation, and another lesser populated more flexible structure [211].

Regarding **computational studies**, URP has not been studied with other computational methods except for the MD and temperature Replica Exchange, which is work performed by us together with collaborators from the University of Portsmouth [212]. In that paper, the UII and URP conformational ensemble obtained by REMD was published. The published REMD results will be presented in the REMD results section here.

8.2 Motivation for our work

Exploring the conformational space of the URP peptide comes as a natural continuation of the work done on UII peptide, since they only differ in the length of the N terminal tail, but have the same ring sequence. Little experimental data is known about the URP peptide, as was the case for UII. Both peptides are known to trigger different biological responses by binding to the same GPCR receptor, so knowing their conformational dynamics, even in the unbound state, may help in understanding their functional diversity.

The URP conformational ensemble was also examined using temperature Replica Exchange and Solute Tempering advanced sampling methods. The performance of the methods will be compared, and then tested against NMR chemical shifts at the end of the chapter.

8.3 REMD simulation

Using the REMD method, three simulation runs were performed using three different starting URP conformations referred to as *Omega I Open, Omega II* and *Lasso.* The starting conformations were obtained from MD simulation from our collaborators. The REMD simulation setup and temperature range used was the same as for the UII peptide.

8.3.1 REMD simulation details

Two simulations were run for 400 ns, while one was run for 300 ns, each using the PMEMD module in AMBER 12 suite programs. The temperature range was generated using the online temperature generator http://folding.bmc.uu.se/remd/ [178] with an overall expected acceptance ratio among replica of 30 % and provided us 64 replicas from 298 K to 543 K. The Amber ff99SB force field was used with explicit TIP3P water model [177]. The initial structures were solvated in a cubic box containing water molecules with periodic boundary conditions and neutralised with 1 Cl^- . The Particle Mesh Ewald [126] was used for the long-range interactions using a 10 Å cutoff. Bonds involving hydrogen were constrained using the SHAKE algorithm [121] with a tolerance of 0.00001 Å. REMD simulations were performed in the NVT ensemble using a Langevin thermostat for the temperature coupling with a collision frequency of 1 ps^{-1}. 200 ps of NVT simulation was used to equilibrate the initial state to the desired temperature for each replica, following a rescaling of the velocities. Using these equilibrated replicas, 400 ns of REMD simulation was performed on each replica, consisting of 25.6 (19.2) μs of molecular dynamics. All exchanges between neighbouring replicas were allowed every 2 ps in the NVT ensemble.

8.3.2 REMD simulation results

Three sets of REMD simulations were performed to explore the URP conformational ensemble. The initial 100 ns of each simulation were taken as equilibration and not included in the simulation analysis. Already established analysis procedures, consisting of analysing β-turn population, hydrogen bond population and cluster state diversity, were also performed for URP peptide. The following sections provide more details.

8.3.2.1 Turn and hydrogen bond populations

Experimental data (Section 8.1) reported the URP conformational flexibility in terms of different β- and γ-turns. Here their population across the ring residues was explored using the definitions given in Section 1.2.1. The populations are given in Table 8.1.

	3,4 type I	3,4 type VIII	4,5 type I	4,5 type II	5,6 type I	5,6 type II	5,6 type VII	4,5,6 γ turn
Omega Open	3.92	1.98	2.35	1.61	14.65	19.05	6.62	5.57
Omega II	3.01	2.11	2.05	3.41	11.86	19.27	7.19	3.67
Lasso	2.16	2.13	1.68	3.36	12.88	17.25	12.44	5.12

Table 8.1: β-turn and γ-turn populations from the three REMD simulations (Omega Open, Omega II and Lasso).

The β-turns centred at residues 5,6 were the most populated during the simulation (Table 8.1) stabilised mostly with the highly populated $Trp^4O - Cys^7H$ hydrogen bond, and other two less populated $Phe^3O - Tyr^6H$ and $Trp^4O - Tyr^6H$ intracyclic hydrogen bonds (Table 8.2).

Hydrogen bond	Omega Open	Omega II	Lasso
$Cys^2O - Lys^5H$	5.08	4.22	3.52
$Cys^2O - Cys^7H$	3.14	3.29	1.62
$Phe^3O - Tyr^6H$	4.17	6.54	5.72
$Phe^3O - Cys^7H$	0.95	0.88	1.12
$Trp^4O - Tyr^6H$	6.77	4.92	7.21
$Trp^4O - Cys^7H$	21.85	17.12	19.23

Table 8.2: Different hydrogen bond populations from the three REMD simulations.

The NMR experiments [51, 211, 214] reported on turns centred at residues 5,6 that is in agreement with the higher population of β-turns centred at these residues in our simulations. Only the $Trp^4O - Tyr^6H$ hydrogen bond was reported in one NMR experiment [214].

8.3.2.2 Torsion based clustering

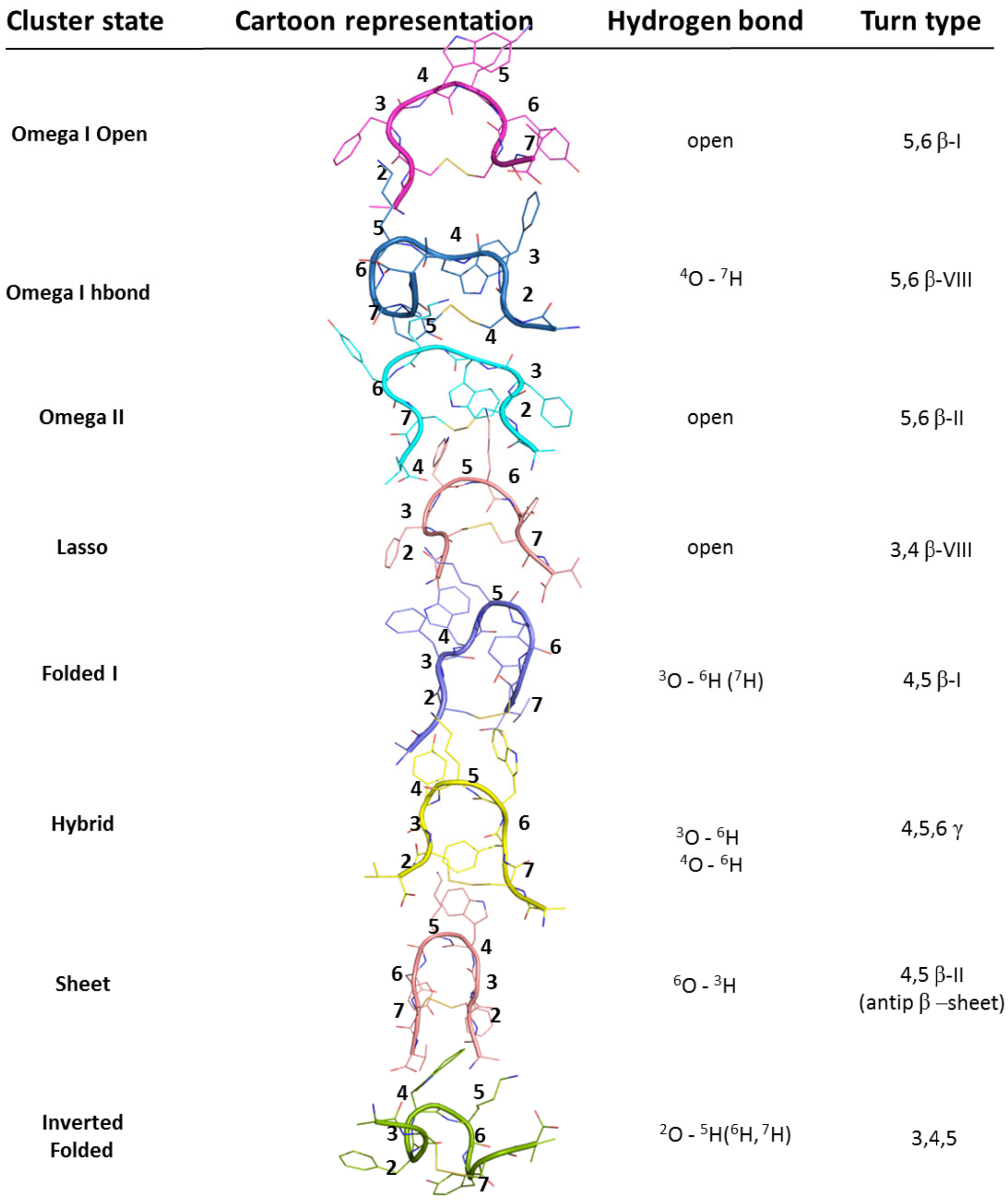
Next, the time series of the ring torsion angles ($\psi 2$, $\phi\psi$ 3-6 and $\phi 7$) was analysed with Dash software to test for the population of the unique URP cluster states. The torsion angles extracted were the same as for UII, to compare their conformations since they share the same six membered ring sequence. Circular similarity analysis of the torsion time trajectory, using the *dashsim* program, revealed that the URP peptide adopts mostly the same ring conformations as UII (Tables B.1, C.1). The URP conformational ensemble was grouped into two major states, *open* and *folded*, further containing a number of substates. An *open* state contains *Omega I Open, Omega I hbond, Omega II* and *Lasso* substates, while *folded* URP cluster state contains *Hybrid, Sheet, Folded I, Folded II, Folded III* and *Inverted Folded* substates (Figure 8.2).

Although torsion based clustering revealed that URP adopts the same ring clustering subtypes as UII peptide, the populations of subtypes *Folded II* and *Folded III* are minor, or not observed in all simulation repeats, so these substates were not included in the final plots. Their population are given in Table 8.3.

Substate	**Omega Open**	**Omega II**	**Lasso**
Folded II	0.0	0.38	0.14
Folded III	0.0	0.34	0.0

Table 8.3: The population of URP *Folded II* and *Folded III* substates in three simulation repeats.

Compared to UII, two new *folded* substates were discriminated in the URP conformational ensemble, referred to as *Hybrid* and *Sheet*. These two states are different in terms of the secondary structure motif, *Hybrid* is described as adopting 4,5,6 γ turn while *Sheet* adopts antiparallel β-sheet. However, the circular similarity between them is 0.71 (Table 8.4), further suggesting that these two states are easily interconverting and can be considered as one, just like *Omega I Open* and *Omega I hbond*, because in the circular similarity analysis, torsion are consider as belonging to the same state if the value of circular similarity is higher than 0.65.



Cluster state	Cartoon representation	Hydrogen bond	Turn type
Omega I Open		open	5,6 β-I
Omega I hbond		^{4}O - ^{7}H	5,6 β-VIII
Omega II		open	5,6 β-II
Lasso		open	3,4 β-VIII
Folded I		^{3}O - ^{6}H (^{7}H)	4,5 β-I
Hybrid		^{3}O - ^{6}H ^{4}O - ^{6}H	4,5,6 γ
Sheet		^{6}O - ^{3}H	4,5 β-II (antip β –sheet)
Inverted Folded		^{2}O - ^{5}H(^{6}H, ^{7}H)	3,4,5

Figure 8.2: The URP cluster substates.

8.3.2.3 Similarity between the UII and URP cluster substates

The circular similarity scores between the UII and URP conformational subtypes show that both peptides adopt almost the same ring subtypes. Table 8.5 gives the circular similarity scores between UII cluster substates and URP cluster substates.

	Omega I Open	**Omega I hbond**	**Omega II**	**Lasso**	**Folded I**	**Hybrid**	**Sheet**	**Inv Folded**
Omega I Open	1.00	0.73	0.53	0.50	0.50	0.55	0.63	0.42
Omega I hbond	0.73	1.00	0.56	0.46	0.58	0.65	0.52	0.37
Omega II	0.53	0.56	1.00	0.54	0.38	0.40	0.41	0.50
Lasso	0.50	0.46	0.54	1.00	0.45	0.31	0.44	0.47
Folded I	0.50	0.58	0.38	0.45	1.00	0.60	0.55	0.43
Hybrid	0.55	0.65	0.40	0.31	0.60	1.00	0.71	0.29
Sheet	0.63	0.52	0.41	0.44	0.55	0.71	1.00	0.32
Inv Folded	0.42	0.37	0.50	0.47	0.43	0.29	0.32	1.00

Table 8.4: The circular similarity between different URP cluster substates.

	Omega I Open	**Omega I hbond**	**Omega II**	**Lasso**	**Folded I**	**Hybrid**	**Sheet**	**Inv Folded**
Omega I Open	0.92	0.95	0.53	0.51	0.50	0.54	0.62	0.38
Omega I hbond	0.95	0.98	0.53	0.46	0.58	0.65	0.52	0.42
Omega II	0.53	0.56	0.98	0.54	0.36	0.40	0.41	0.49
Folded I	0.50	0.58	0.38	0.45	0.98	0.60	0.55	0.43
Folded II	0.49	0.40	0.43	0.38	0.75	0.42	0.45	0.41
Folded III	0.46	0.42	0.41	0.42	0.56	0.42	0.34	0.47
Folded IVb2	0.62	0.49	0.43	0.34	0.61	0.89	0.66	0.62
Inv Folded	0.42	0.37	0.50	0.47	0.43	0.29	0.32	0.99
Lasso	0.46	0.50	0.54	0.98	0.45	0.31	0.44	0.47
Circle	0.66	0.65	0.42	0.65	0.53	0.36	0.51	0.46
Scoop	0.53	0.59	0.39	0.54	0.49	0.53	0.61	0.41

Table 8.5: The circular similarity between different UII (far left column) and the most populated URP cluster substates.

8.3.3 The time evolution of the cluster substates

Having defined the URP cluster ensemble, we also wanted to check for the URP substates time evolution, as given in Figure 8.3. It shows that the most populated substate is *Omega II*, followed by *Open Omega*. The only exception is the REMD Omega II ensemble where there is a higher population of Omega II substate, but this could be due to the incompletely converged simulation since it was also the starting conformation in this simulation repeat. Other *open* and *folded* substates are minorly populated, and their populations differ only in a few percentages between the REMD simulation repeats.

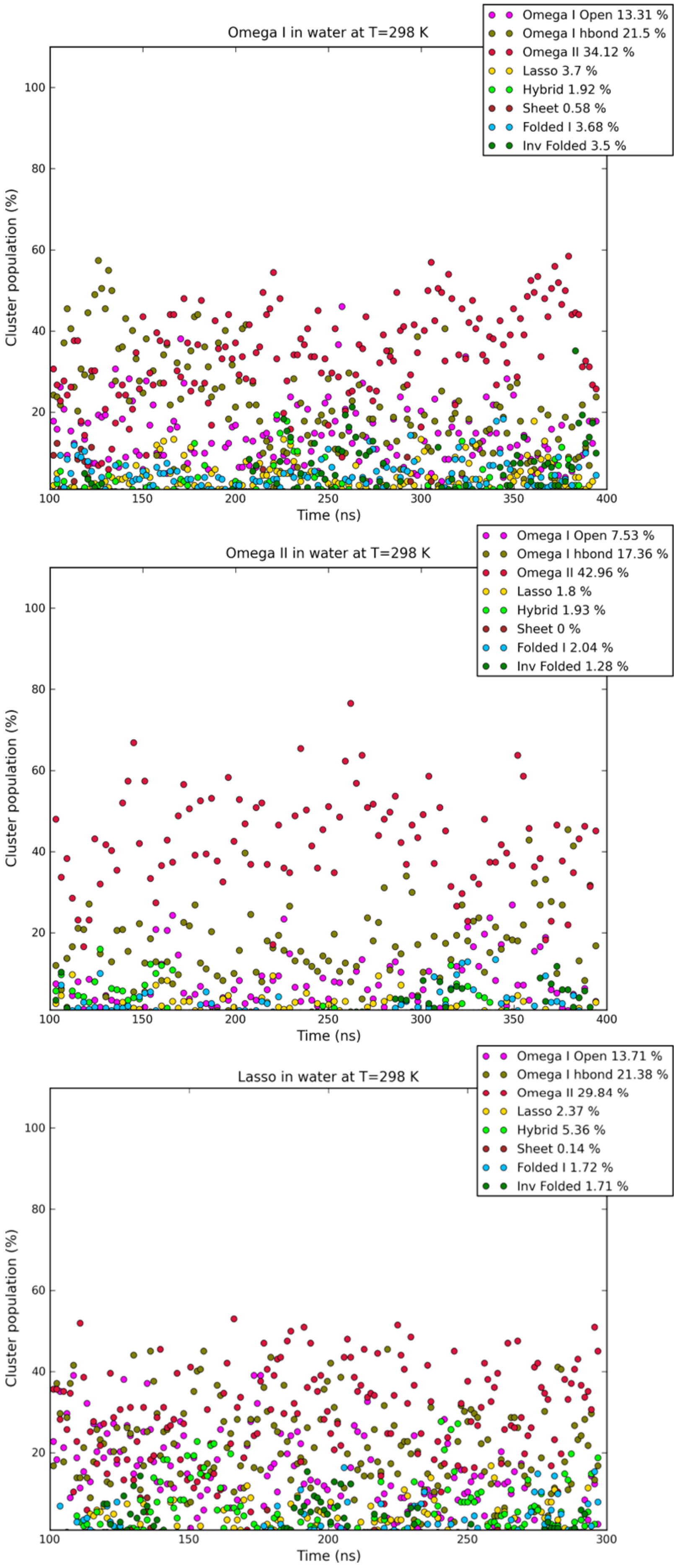


Figure 8.3: The time distribution of the URP cluster substates in the three REMD ensembles.

8.4 REST simulations

The URP conformational ensemble was also explored with the Solute Tempering method. The efficiency of this method was shown in the example of the structurally more complex Urotensin UII peptide, so here it was applied too. The method was run with three different URP starting conformation, the same as for with REMD (*Open I Omega, Omega II* and *Lasso*).

8.4.1 REST simulation details

Three URP cluster substates were used to perform REST simulations. The method was run in Gromacs with the Plumed patch [155]. The Amber ff99SB, was used with explicit TIP3P water model [177]. The system was neutralised with the Cl^- counterion. The Particle Mesh Ewald [126] was used for the long-range interactions using a 10 Å cutoff. Bonds involving hydrogen were constrained using the SHAKE algorithm [121] with a tolerance of 0.00001 Å. REST simulations were performed in the NVT ensemble using a Langevin thermostat for the temperature coupling with a collision frequency of 1 ps^{-1}. The simulations were run for 300 ns using 12 replicas in the temperature range 298 K - 900 K. The replicas were geometrically distributed to give the acceptance ratio between 20 and 35 %.

8.4.2 REST simulation results

8.4.2.1 β-turn and hydrogen bond population

The hydrogen bond and β-turn analysis was performed in accordance with the already established procedure to test the peptide structural diversity independent of the torsion based cluster analysis.

The populations of different intra-ring β-turns was explored, and in Table 8.6 the most populated β-turns are given, together with the population of the experimentally reported γ turn. A Table 8.6 shows that the URP peptide prefers ring conformations with β-turns centred at residues Lys^5O and Tyr^6H, followed by a highly populated 4,5 β-turn.

Table 8.7 shows that β-turns are often stabilised by the $Phe^3O - Tyr^6H$, and Trp^4O with residues Tyr^6H and Cys^7H, hydrogen bonds.

	3,4 type I	**3,4 type VIII**	**4,5 type I**	**4,5 type II**	**5,6 type I**	**5,6 type II**	**5,6 type VIII**	**4,5,6 γ turn**
Omega Open	0.35	1.95	6.71	2.91	24.26	4.85	5.87	2.67
Omega II	0.63	2.19	5.72	2.38	29.12	2.86	6.61	4.57
Lasso	0.29	1.97	3.15	2.25	20.65	3.88	8.09	4.69

Table 8.6: β-turn type populations from the three REST simulations

Hydrogen bond	**Omega Open**	**Omega II**	**Lasso**
$Cys^2O - Lys^5H$	2.51	3.77	2.13
$Cys^2O - Cys^7H$	1.98	3.64	4.29
$Phe^3O - Tyr^6H$	7.56	5.24	6.06
$Phe^3O - Cys^7H$	1.39	3.15	3.05
$Trp^4O - Tyr^6H$	6.09	8.29	6.72
$Trp^4O - Cys^7H$	16.41	16.82	18.84

Table 8.7: Different hydrogen bond populations from the three REST simulations.

8.4.3 The time evolution of the cluster substates

The URP cluster ensemble was determined in the same way as was done using temperature Replica Exchange (Section 8.3.2.2). The identified URP substates were also plotted during the simulation to check for the time evolution of the substates, and then compared with the REMD simulation runs.

Since the *folded* substates *Folded II* and *Folded III* were almost not populated in the REMD simulation repeats, and are taken as transient states, their populations are also only given in the Table 8.8 for Solute Tempering simulation repeats. As it can be observed, they are not present or observed in only a few frames during the REST simulations.

Substate	**Omega Open**	**Omega II**	**Lasso**
Folded II	0.0	0.0	0.12
Folded III	0.90	0.0	0.00

Table 8.8: The population of URP *Folded II* and *Folded III* substates in three simulation repeats.

The final URP cluster state populations across the REST simulations are given in Figure 8.4 showing that URP is exchanging between substates regularly.

Overall, URP prefers *open* clusters states characterised by β-turns centred at residues 5,6 or 3,4. In particular, the substate *Omega I Open* and *hbond* are populated almost one third of the simulation time. The next most populated substate also belongs to the *open* cluster group, *Omega II* which is populated

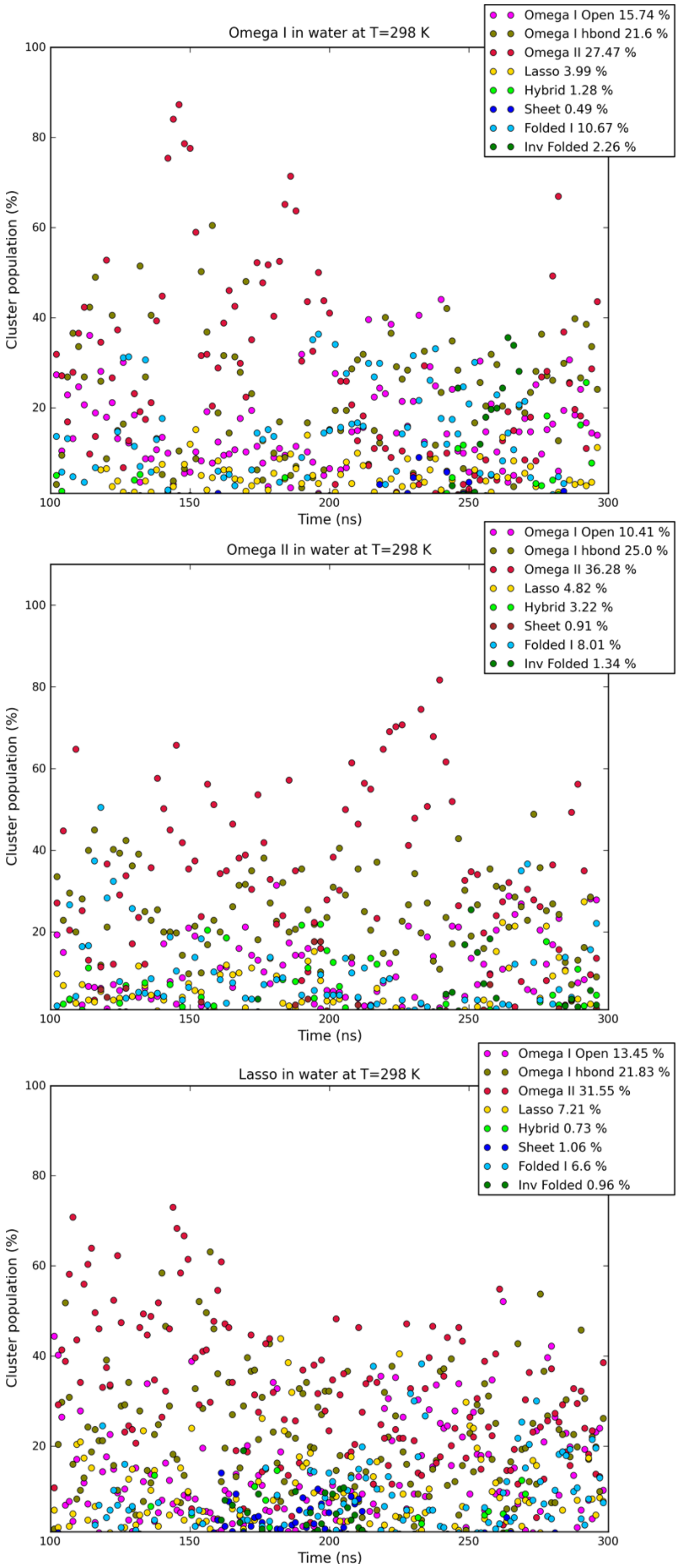


Figure 8.4: The time distribution of the URP cluster substates in the three Solute Tempering ensembles.

between 30 % and 35 % during the simulation time. The most populated *folded* substate is *Folded I*, 5 % to 10 %. This substate is characterised by a 4,5 β-turn, also populated approximately 5 - 8 %, further confirming the population of this cluster substate. The URP structure found in SDS micelles was described as being 4,5 β-turn as well [211] (see Section 8.1). Finally, two URP specific substates *Hybrid* and *Sheet* are together populated for 2 - 4 % depending on the simulation repeat.

8.5 URP chemical shifts

After the analysis of the URP conformational ensemble, the data were validated against the proton chemical shifts obtained at pH 6.0 and a temperature of 298 K [212]. The proton chemical shifts were calculated using Gaussian09 software [116] with B3LYP/6-31G(d) level of DFT theory. The procedure applied for URP cluster representatives is the same as already described in Sections 4.7.0.2 for AVP and then applied to all peptides studied in this work.

8.5.0.1 The choice of the representative structures

The representative structures were chosen to fulfil the same conditions as for all already studied peptide; to be scattered in the approximately equal intervals along the trajectory, and to be within the 1 SD of the torsion angle distribution B.1. There are six representative structures for each URP substate, and the variance between the chemical shift types for each cluster state is given in Figure 8.5. It shows that URP cluster states adopt very tight chemical shift ranges with the exception of Omega Open Lys^8 HA, HB2, HB3, HG2 chemical shift types.

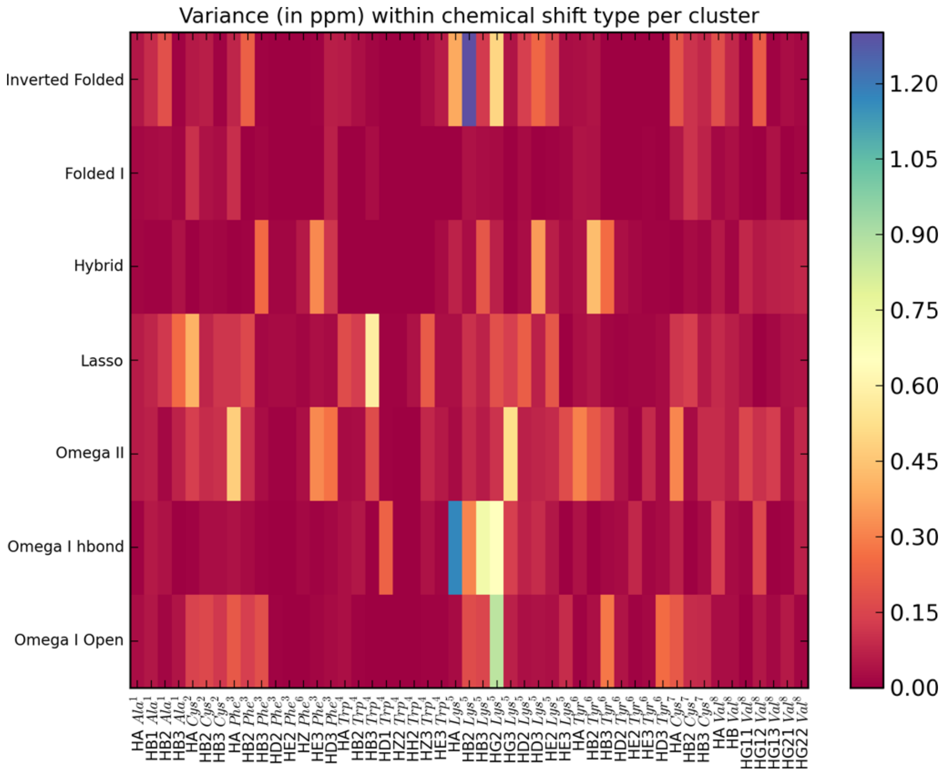


Figure 8.5: The variance within each chemical shift type for the given number of substate representative structure.

Next, the statistical analysis was used to check on the peptide substate variance. Three statistical measures of similarity MUE, MSE and R^2 were used. Figure 8.6 shows the R^2 distribution of the values for individual substate members compared against experimental data, while Figure 8.7 shows their bootstrapped R^2 values.

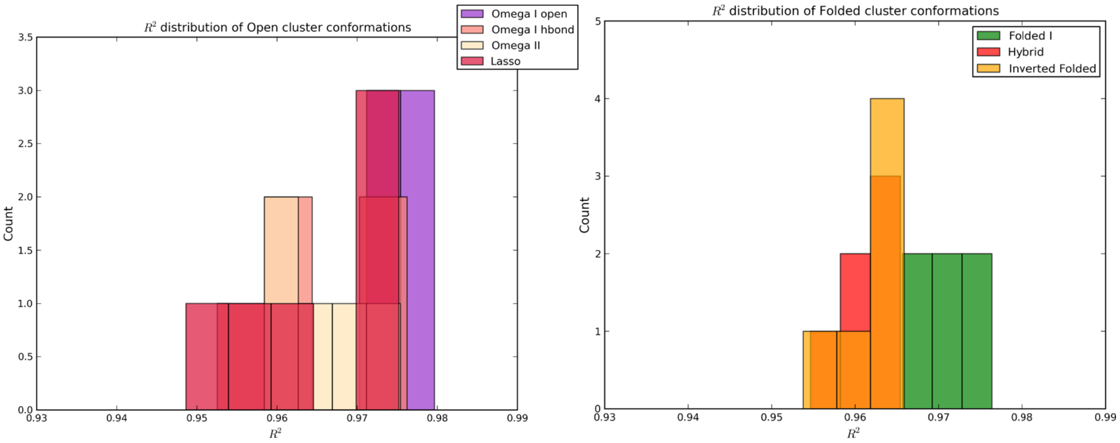


Figure 8.6: The R^2 distribution of the values for individual substate structure chemical shifts compared against experimental data.

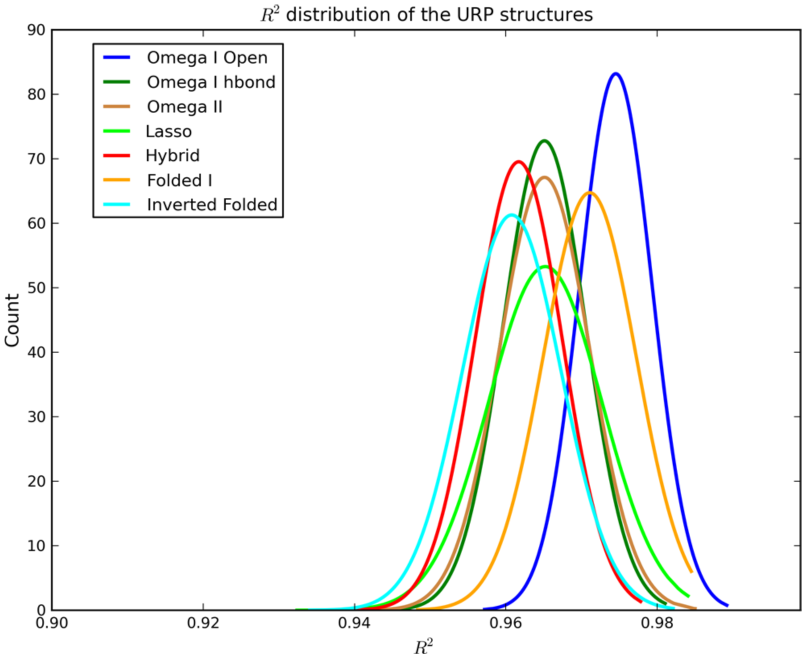


Figure 8.7: The bootstrapped R^2 distribution of the URP cluster substates.

Figure 8.7 shows that URP substates have bootstrapped R^2 mean values in the range 0.96 to 0.98, with the highest value for *Omega I Open*, followed by *Folded I* substate. Bootstrapped MUE values also show the smallest value for *Omega I Open* and *Folded I* substates (Table 8.9).

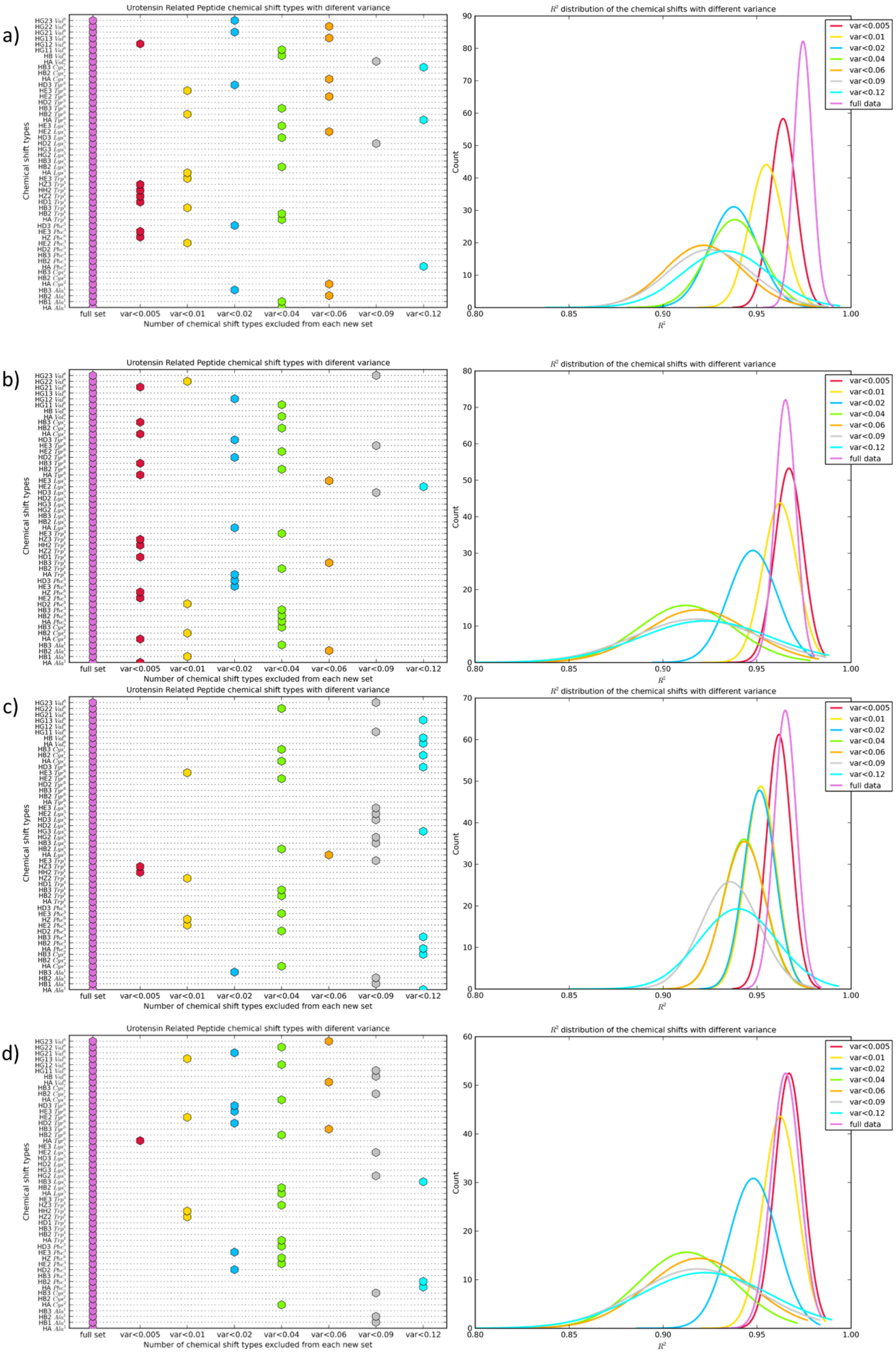
	MUE	**MSE**	**R^2**
Omega I Open	$0.205 < 0.266 <0.326$	$0.028 < 0.118 < 0.208$	$0.965< 0.974 < 0.983$
Omega I hbond	$0.275< 0.355 < 0.436$	$-0.128 < 0.002 <0.128$	$0.954 < 0.965 < 0.975$
Omega II	$0.263 < 0.343 < 0.423$	$-0.044 < 0.078 < 0.201$	$0.953 < 0.965< 0.976$
Lasso	$0.279 < 0.364< 0.449$	$-0.101 < 0.031 < 0.164$	$0.950 < 0.965 < 0.980$
Hybrid	$0.262 < 0.353 < 0.444$	$-0.121< 0.012 < 0.147$	$0.950 < 0.961 < 0.972$
Folded I	$0.230 < 0.298< 0.367$	$-0.019 < 0.086 < 0.191$	$0.958 < 0.971 < 0.983$
Inv Folded	$0.255 < 0.352 <0.448$	$-0.116 < 0.021< 0.158$	$0.948 < 0.960 < 0.973$

Table 8.9: The bootstrapped values of the three statistical measures of similarity, Mean Unsigned Error (MUE), Mean Signed Error (MSE) and coefficient of determination (R^2) for different URP cluster members

8.5.0.2 Analysis of the individual chemical shift types

The idea of this analysis to check how cluster state distribution depends on the individual chemical shift types is given in Figure 8.8.

Compared to other peptides, URP chemical shifts adopt very good R^2 values for all cluster states but *Inverted Folded* (Figure 8.8 e)). In this case, the correlation coefficient drops significantly for the chemical shift types with variance higher than 0.04 ppm. Other cluster states, *Omega Open, Omega II, Lasso* and *Folded I* show that almost all chemical shift types contribute similarly to the final R^2 distribution.


a)
b)
c)
d)
Urotensin Related Peptide chemical shift types with diferent variance
Chemical shift types
HG23 Val⁸
HG22 Val⁸
HG21 Val⁸
HG13 Val⁸
HG12 Val⁸
HG11 Val⁸
HB Val⁸
HA Val⁸
HB3 Cys⁷
HB2 Cys⁷
HA Cys⁷
HD3 Tyr⁶
HE3 Tyr⁶
HE2 Tyr⁶
HD2 Tyr⁶
HB3 Tyr⁶
HB2 Tyr⁶
HA Tyr⁶
HE3 Lys⁵
HE2 Lys⁵
HD3 Lys⁵
HD2 Lys⁵
HG3 Lys⁵
HG2 Lys⁵
HB3 Lys⁵
HB2 Lys⁵
HA Lys⁵
HE3 Trp⁴
HZ3 Trp⁴
HH2 Trp⁴
HZ2 Trp⁴
HD1 Trp⁴
HB3 Trp⁴
HB2 Trp⁴
HA Trp⁴
HD3 Phe³
HE3 Phe³
HZ Phe³
HE2 Phe³
HD2 Phe³
HB3 Phe³
HB2 Phe³
HA Phe³
HB3 Cys²
HB2 Cys²
HA Cys²
HB3 Ala¹
HB2 Ala¹
HB1 Ala¹
HA Ala¹
full set
var<0.005
var<0.01
var<0.02
var<0.04
var<0.06
var<0.09
var<0.12
Number of chemical shift types excluded from each new set
R² distribution of the chemical shifts with different variance
Count
R²
0.80
0.85
0.90
0.95
1.00
0
10
20
30
40
50
60
70
80
90
var<0.005
var<0.01
var<0.02
var<0.04
var<0.06
var<0.09
var<0.12
full data

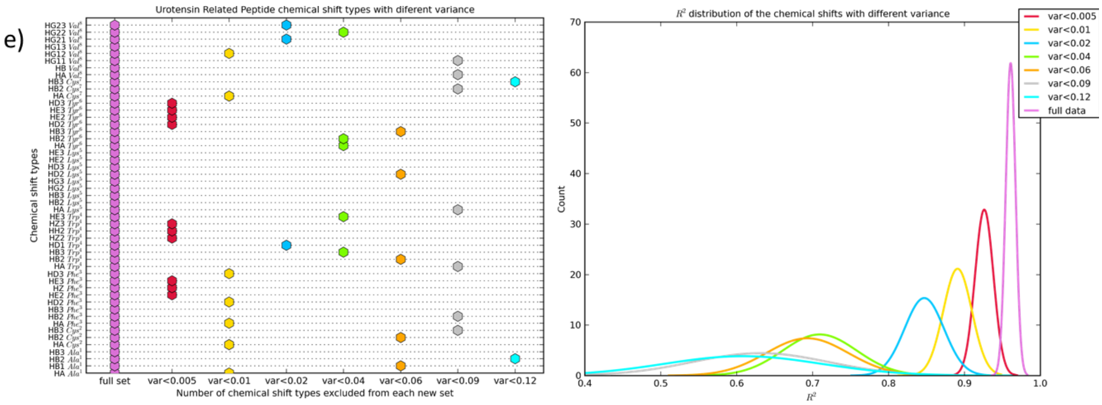


Figure 8.8: A dependence of the cluster state distribution on the chemical shift data sets. The chemical shifts were extracted from the initial data sets depending on their variance (left figures), and then compared with the matching experimental chemical shift data (right figures). The plots correspond to a) Open Omega , b) Omega II, c) Lasso, d) Folded I, e) Inverted Folded cluster substates.

8.6 Ensemble model

The concept of the ensemble model was finally applied to the URP peptide as well. Figure 8.9 gives the bootstrapped R^2 values for six different ensemble models obtained as a weighted sum of chemical shifts with simulation populations.

Figure 8.9 shows that all simulation ensembles, except that derived from the REMD Omega II starting structure, have very tight mean R^2 values, approximately in the range from 0.983 to 0.986. Of all the simulation repeats, the best agreement with experimental data is for the Solute Tempering Omega I ensemble (shown in orange in Figure 8.9). This ensemble model is closely followed by other Solute Tempering and Replica Exchange simulations, with the exception of the REMD Omega II simulation.

Next, a comparison between the bootstrapped R^2 range for the individual cluster members (Figure 8.7) and ensemble models (Figure 8.9) shows that the ensemble model is outperforming any single cluster state. Table 8.10 gives the mean bootstrapped R^2 values for the easier comparison, showing that the highest R^2 value of the best single state *Omega I Open* (0.9745) is still lower that the worst performing ensemble REMD Omega II simulation (0.9804).

Then, we can also compare the populations of the cluster substates in the simulations, with the performance of ensemble model. The values of the populations follow, giving the best performing ensemble model in bold:

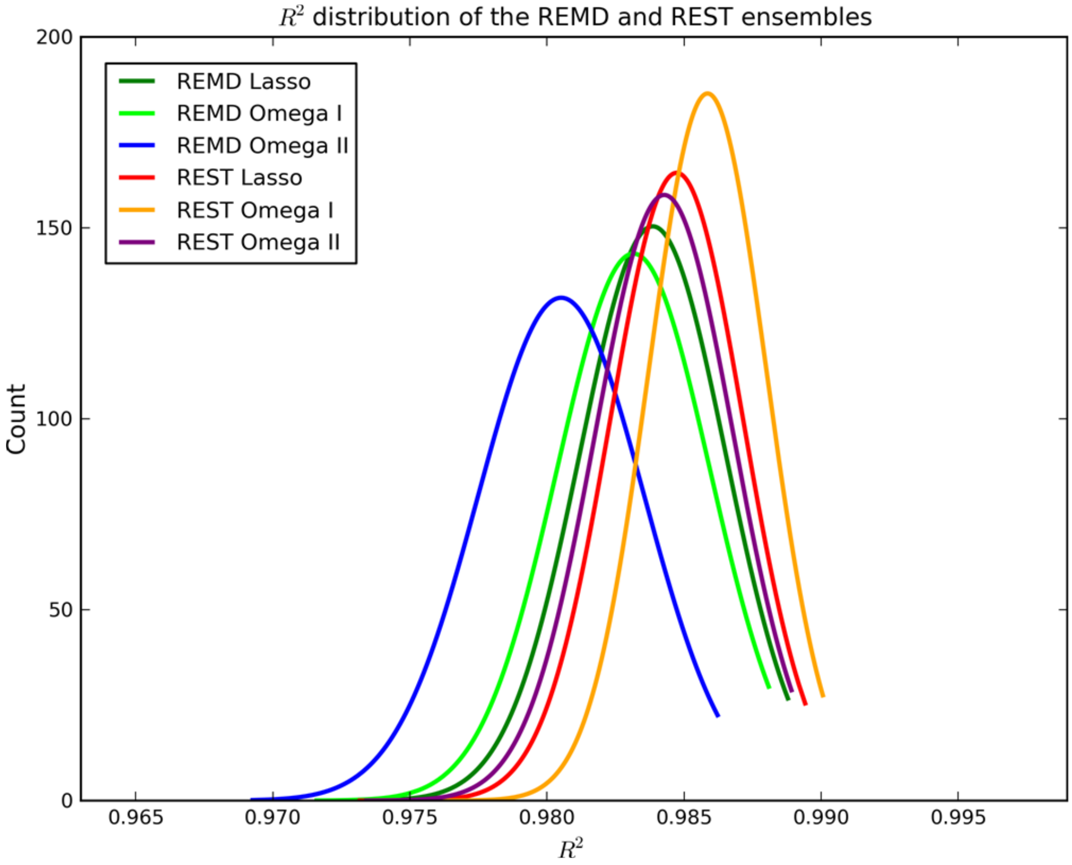


Figure 8.9: The bootstrapped R^2 distribution of the URP simulation repeats.

Cluster substate	R^2 value	**Ensemble model**	R^2 value
Omega I Open	0.9745	**REMD Omega I**	0.9830
Omega I hbond	0.9651	**REMD Omega II**	0.9804
Omega II	0.9650	**REMD Lasso**	0.9837
Lasso	0.9651	**REST Omega I**	0.9858
Hybrid	0.9617	**REST Omega II**	0.9842
Folded I	0.9710	**REST Lasso**	0.9847
Inverted Folded	0.9608		

Table 8.10: A list of the mean boostrapped R^2 values for the individual URP cluster substates and ensemble models

- **REST Omega I**: Omega Open - 38 % : Omega II - 28 %: Lasso - 7 % : Folded I - 10 %
- REST Lasso: Omega Open - 36 % : Omega II - 32 % : Lasso - 4 % : Folded I - 6 %
- REST Omega II: Omega Open - 36 % : Omega II - 36 %: Lasso - 5 % : Folded I - 8 %

The Solute Tempering Omega I ensemble model has slightly higher bootstrapped R^2 value than the other ensembles (Table 8.10). The population of the individual substates in this ensemble is such that there is a higher percentage of the *Omega Open* substate, a lower population of *Omega II* substate and higher population of *Folded I* substate, compared to other ensembles.

However, the difference in the population between ensemble substates is very small, no more than 5 %, suggesting that it is the overall ensemble model with the combination of substate ratio *Open Omega* 35 - 40 %, *Omega II* 30 -35 %, *Lasso* 5 - 10 %, and *Folded I* 5 - 10 % that outperforms the individual URP substates.

8.6.0.1 Optimal population ratios

Figure 8.10 shows the URP ensemble population ratios which have the best agreement with experimental chemical shift data ($R^2 > 0.99$). As can be observed, different population ratios may give the same result, with *Omega Open* and *Folded I* adopting the highest populations in all obtained optimal population ratios, which was also observed in the simulation data. These results supports the idea that the simulation conformation ratios are converged, and have better agreement with experimental data compared to single cluster conformations.

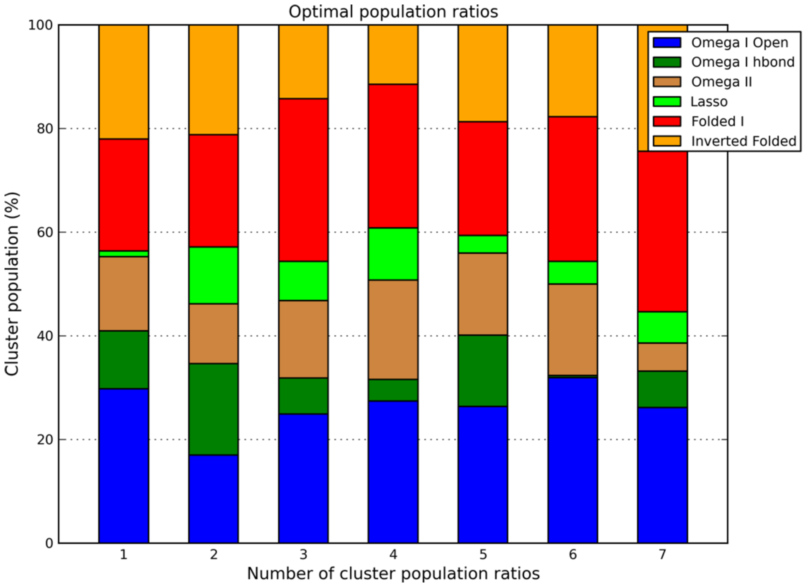


Figure 8.10: The optimal population ratio between URP cluster substates which yields the best agreement with experimental chemical shift data.

8.7 REMD vs. REST conformational ensemble

This Section gives an overview of the URP conformational ensemble obtained by temperature Replica Exchange and Solute Tempering. The quantitative description of the substate population ratios is plotted in Figure 8.11.

The URP conformational ensembles was grouped into two major conformational states, *open* and *folded*. In total, the *open* state is more populated than the *folded* state, the *Omega I Open*, *Omega I hbond* and *Omega II open* substates make 60 - 70 % of the total URP conformational ensembles in both the REMD and REST simulation ensembles.

The analysis of the individual *open* substates shows that the *Open Omega* substates (*Omega I Open* and *Omega I hbond*) are populated between 35 - 40 % and *Omega II* is between 30 and 35 % in both REMD and REST.

The next most populated substate belongs to *folded* group, and is referred to as *Folded I*. This substate, characterised by a turn centred at residues 5,6, is more populated in Solute Tempering than in the REMD method. The same was observed for the UII peptide, suggesting that to see the *folded* state in REMD, longer simulation time is required or the use of higher temperatures in the temperature space.

The *Folded I* state is interesting because the experimental data report on the folded-like structure with β-turn centred at residue 4,5, and on the high similarity

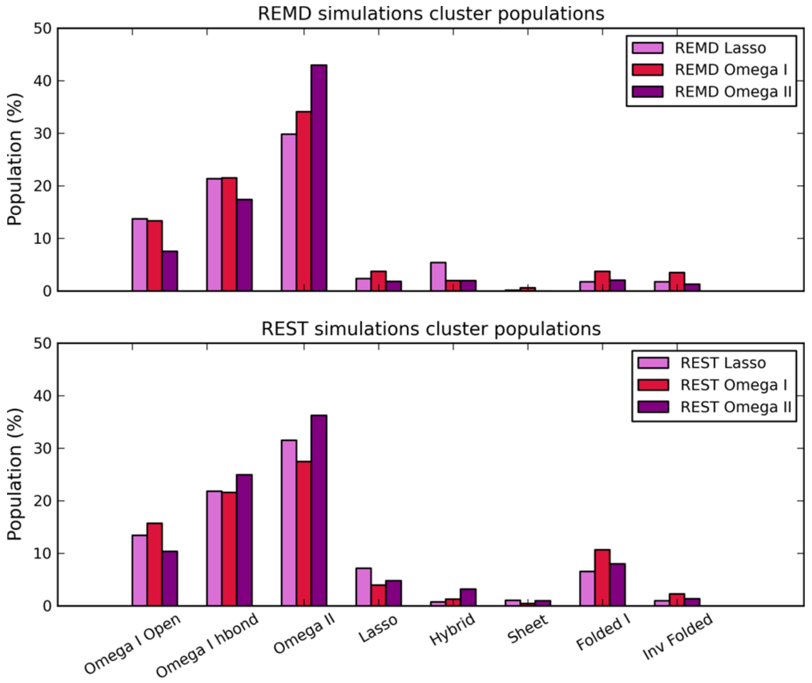


Figure 8.11: The comparison between the substate populations between REMD and REST simulation ensembles.

of URP and UII ring conformations. Moreover, the chemical shift analysis revealed that the simulation ensembles with the best agreement with experimental data suggest high population of *Open Omega* substates with higher population of *Folded I* substate.

Other substates observed in the simulations are *Lasso, Hybrid, Sheet* and *Inverted Folded*, which are all similarly populated across all simulation repeats. The substate populations differ no more than 5 % across the method repeats. The MD simulations performed by Haensele et al [212] did not observed all UII substates in the URP conformational ensemble, such as *Folded I, Folded II, Folded III* and *Lasso*. Therefore, these MD simulations can not be considered as converged.

8.8 Conclusions

In this chapter, the published results of the temperature Replica Exchange simulations were compared against the Solute Tempering simulation runs. The data show that URP is mostly found in the *open* conformational substates, named as *Open Omega* and *Omega II*, characterised by turn types centred at residues 5,6. The populations of the overall *open* states is approximately 60 - 70 %. However, in the Solute Tempering simulation runs, it was observed a bit higher population of

Folded I substate whose structural characterisation is similar to the Brancaccio et al. [51] structure description observed in experiment with SDS micelles. Moreover, the bootstrapped R^2 value of the *Folded I* representative structure chemical shifts also show high values compared to other URP cluster substates. Other identified substates had minor populations.

Chapter 9

Comparison between UII and URP ensembles

The conformational ensemble of UII and URP hormone peptides was examined using two enhanced sampling methods, temperature Replica Exchange and Solute Tempering. The peptides conformational convergence was extensively studied by doing three simulation repeats per method, each starting with different geometry.

A temperature Replica Exchange simulations were run for 25 μs for URP peptide, and 32 μs for UII peptide per simulation. All the Solute Tempering simulations were run for 300 ns per replica, giving in total 3.6 μs simulation time for each simulation repeat.

Both peptides were revealed to adopt the same dominant classes of conformations, *open* and *folded* (Section 8.3.2.2, Table 8.5), but they differ in the populations of conformational subtypes (Tables 9.1, 9.2).

Furthermore, the NMR ensemble model was applied to test the idea of peptide conformational flexibility, confirming the simulation predictions of peptides existing in an ensemble of interchanging conformations, rather than a single conformation.

In terms of structural diversity, the peptides share both *open* and *folded* cluster states. *Open* states observed in both peptides are *Omega I Open, Omega I hbond, Omega II* and *Lasso.* Two UII *open* substates not observed in URP ensembles are *Scoop* and *Circle. Scoop* can be considered as transient state because it is not populated more than 1 % in any of the simulation repeats, while *Circle* is very

Sampling method	Omega Open	Omega II	Folded I	Folded II	Folded III	Folded IVb2	Inv. Folded	Lasso	Scoop	Circle
UII REMD	6.33 ± 2.71	2.63±1.17	3.13±1.71	2.79±1.17	3.68±3.5	0.33±0.11	9.71±2.94	44.95±3.19	0.13±0.09	1.45±0.31
UII REST	7.59±2.95	5.03±1.33	7.88±1.91	2.06±1.47	0.99±0.44	0.79±0.08	7.78±2.71	43.35±1.42	0.22±0.13	2.44±1.12

Table 9.1: The mean and SD of the UII cluster substates averaged over three ensembles per enhanced sampling method.

Sampling method	Omega Open	Omega II	Folded I	Inv. Folded	Lasso	Circle	Hybrid	Sheet
URP REMD	15.84±2.21	35.64±5.46	2.48±0.85	2.16±0.96	2.62±0.79	3.07±1.61	0.23±0.24	0.00
URP REST	18.42±1.86	31.76±3.56	8.32±1.78	1.74±1.06	5.31±1.36	1.74±1.06	0.84±0.25	0.85±0.24

Table 9.2: The mean and SD of the URP cluster substates averaged over three ensembles per enhanced sampling method. The URP conformational substates not given in the Table, *Folded II, Folded III, Scoop* and *Circle* were very little or not populated in the all enhanced sampling ensembles.

similar to *Omega I Open* state (circular similarity = 0.66).

A UII *folded* cluster ensemble contains five cluster substates, all shared with URP. However, two of them *Folded II* and *Folded III* were almost not observed in the URP simulations, and are considered as transient states. A UII *Folded IVb2* conformation is 89 % similar to URP *Hybrid* substate, revealed by circular similarity analysis (Table 8.5).

The two most populated UII *folded* substates *Folded I* and *Inverted Folded* are conformational substates shared with URP. A *Folded I* substate was observed to be somewhat more populated in Solute Tempering simulations than in temperature Replica Exchange for both peptides (Tables 9.1, 9.2), suggesting that higher energy is maybe needed to observe convergence of this *folded* state, which was characterised to be stabilised with $Phe^3O - Tyr^6H$ and $Phe^3O - Cys^7H$ hydrogen bonds.

To summarise, from the overview of the conformational substate populations between UII and URP, it can be seen that all UII major populated states are also observed in the URP conformational ensemble. The UII conformational ensemble seems to be more diverse between the substate populations, and this could be due to the long N terminal tail missing in the URP structure.

Regarding the population of the major cluster substates, the UII and URP cluster substates differ. The result obtained for UII suggests that conformational substate named *Lasso* was the most populated substate for UII, approximately 45 % (Table 9.1). Other highly populated substates in both simulations were *Omega Open* and *Folded I* (Table 9.1. If compared to URP conformational ensemble, then it can be observed that the URP preferred substate is *Omega Open* followed by *Omega II.* The data suggest that the peptides have preferences for different dominant ring conformers, despite the same ring sequence. Moreover, the conformational equilibrium is shifted towards *open* conformation class compared to *folded* in the approximative ratio 3:1 for both peptides.

Since it is known that the UII and URP peptides by binding to the same receptor trigger different biological responses, their structural diversity may help in gaining an insight into receptor activation mechanisms. More about peptides biological function is given in the next section.

9.1 Biological activity

Urotensin II and Urotensin Related Peptide are two hormone peptides that exert a variety of physiological roles in our body. Both peptides were first discovered in the positions encoding for the motoneurons and spinal cord in the human genome. However, mRNAs encoding the peptides were then also found in peripheral tissues such as heart, spleen, kidney, prostate, pituitary - just to name few [215].

Both peptides are endogenous ligands to the same GPCR receptor, initially identified as a human analogue of the GPR14 receptor [205], but then renamed to Urotensin II receptor (UTR) [216]. Just like UII and URP, UTR is also widely expressed in the central nervous system, and to a lesser extent in peripheral tissues. Together, the peptides with the receptor are referred to as *urotensinergic system* which has an important cardiovascular role as well as endocrine and behavioural effects [50]. In particular, UII has both vasoconstriction and vasodilutive roles in our body [217, 218].

9.1.1 Urotensin II receptor

9.1.1.1 Structure

The Urotensin II receptor (GPR14, UTR) belongs to a class A rhodopsin family G protein coupled receptors. It has a 389 long amino acid sequence organized into 7 transmembrane (TM) helices, a common structural GPCR motif, connected by several extracellular and intracellular loops. Although the UTR receptor is conserved across the species, in terms of sequence similarity, rat UTR which contains 386 amino acids, shows only 75 % similarity with the human UTR, while the sequence of human and monkey UTR is almost identical [219].

9.1.1.2 Signalling cascade

The urotosingeneric system is involved in the number of cascade pathways which strongly depend on the tissue in which the UTR is expressed.

In the course of cascade events, upon binding of the ligand, UTR interacts with the G Protein alpha subunit, Gαq11 involved in activating Protein Kinase C (PKC). This then activates phospholipase C which increases the intracellular amount of calcium through the activation of IP3 which is an intracellular molecule that acts as a secondary messenger. IP3 will then release calcium which then activates PKC. If the UTR receptor is situated on the nuclear membrane, then calcium ions are known to be involved in the regulation of the gene transcription [220].

9.1.1.3 Receptor expression and binding

Furthermore, UTR was also discovered to exist on nuclear membrane only recently where, besides being responsible for the regulation of gene transcription, it is also included in ionic homoeostasis, cellular proliferation, and remodelling. However, the orientation of the receptor remains to be unclear. It is believed that the ligand binding site is situated either within or outside the nucleus. The active site is at the C terminal part of the receptor so the signal could be sent either from the nucleus to the cytosol or the other way round [215, 220].

Moreover, the receptor activation mechanism remains unclear. Although it was suggested that UII and URP peptides bind to the same binding site on the receptor [221], their signalling cascade differs [222], suggesting that the peptides may activate the UTR receptor in a different manner.

One of the suggested receptor activation mechanisms includes the idea that UTR can discriminate between UII and URP conformations, what is known as *biased agonism* [222]. Two modes of actions were proposed, one hypothesis is that UTR can discriminate between the cyclic parts of the peptides [222] [50], while other suggests the idea that it is the N terminal tail of the UII makes the difference in the receptor recognition [51].

Nevertheless, the conformational ensemble obtained using enhanced sampling methods can contribute to the hypothesised mechanisms because it is believed that it contains complete UII and URP conformational ensemble, and it will likely contain the peptide's bioactive conformation. It also revealed that UII and URP peptides adopt the same ring conformational subtypes, suggesting that it may not be the ring part of the UII and URP peptides that triggers the activation of secondary messengers, but the UII N-terminal interactions with UTR receptor. This activation mechanism was also suggested by Brancaccio et al.[51]. However, a difference in receptor recognition could also make longer UII N terminal tail.

Chapter 10

Comparison between peptide ensembles

Molecular dynamics simulations have been widely used to provide atomistic details of the conformational changes in peptides. However, their accuracy is limited by the long time scales required to see many conformational changes between the peptide conformational states. For example, Haensele et al. [212] performed a number of 5 μs long simulations, without observing a change in the peptide's conformational state from the starting structure conformation.

To address the problem of the slow sampling, different enhanced sampling methods have been developed, some of them reviewed in the chapter 3.4. Two of them, temperature Replica Exchange and Solute Tempering, were applied in this work to examine the conformational dynamics of the four cyclic hormone peptides in solution.

The use of computational methods, in particular enhanced sampling methods, to study the conformational changes of the peptides is especially advantageous if there is little experimental data known about their structure or if their crystal structure is unknown, but the peptides have been found to have different important biological functions.

However, to validate the approach, first the enhanced sampling method limitations and required simulation time to see converged sampling need to be assessed for the peptides of similar size, structural characteristics and known crystal structures.

In our work, the conformational sampling of four cyclic peptides was examined, with two of them, AVP and OT, with known crystal structures. The approach taken with these peptides was further applied to another two cyclic peptides, UII and URP, which have very little experimental data reported about their structural diversity, but are widely expressed in our body and involved in multiple pathophysiological processes [50].

Therefore, the idea was to examine the peptide conformational ensembles using advanced sampling approaches in combination with NMR chemical shifts data in order to gain an insight into the peptide's structural ensemble, and in turn to connect these with their functional properties.

Vasopressin and Oxytocin are both nine amino acids long peptides differing only in two residues, one in the ring part of the structure, and another in the tail part. To investigate the extent to which a difference in their sequence may contribute to their functional diversity, their conformational ensembles were explored and compared against known experimental data.

The crystal structures of AVP (PDB:1JK4) and OT (PDB:1NPO) reported the same *folded* ring conformations in two independent experiments. An additional AVP crystal structure (PDB:1YF4) reported a more *open* ring conformation, while for OT another *folded*-like crystal structure was determined (PDB:1XY1). The same *folded* ring conformations for both peptides (PDB IDs: 1JK4, 1NPO) were co-crystallised in the complex with the same binding partner - neurophysin, while the *open* AVP conformation was co-crystallised with a different binding partner - trypsin, suggesting that the peptides can adopt different bound conformations depending on the binding partner.

The structural ensembles of AVP and OT obtained by running long time scale enhanced sampling methods revealed that both peptides adopt the same ring conformations with the same dominant conformers, but in a different population ratio. The most populated conformational state in both peptides obtained by running the simulations are the *folded* crystal structures (PDB IDs: 1JK4, 1NPO). In addition, OT was found as almost entirely adopting this crystal structure, observed 70 - 80 % during the simulation time, while AVP showed more conformational flexibility, with the folded conformations observed between 30 to 40 % of the simulation time.

The next most populated conformational state in OT was another crystallographic determined OT conformation, named as *Twisted Saddle* and populated

10 %, while for AVP other identified conformational substates were similarly populated, 10 to 15 % for *Open, Clinched Open* and *Twisted Saddle* cluster states. Therefore, these data suggest that the bound peptide conformation can be found as a dominant conformational state in the limit of the converged simulation runs. The population ratios between the AVP and OT conformational ensembles differs, which could be linked to different binding affinity of the peptides to the receptors. While AVP binds with equal affinity to the OT receptor OTR and AVP receptor V2R, Oxytocin was not very active at AVP receptors, suggesting that the difference in their sequence could affect the preferred structural state responsible for their biological actions.

After the AVP and OT conformational ensembles showed very good agreement with structural experimental data, the next step in validating the simulation predictions was to examine them against NMR chemical shifts. The ensemble model approach, consisting of weighting the chemical shifts values with the ensemble populations, was used. Each peptide's cluster substate was represented by a few structures whose chemical shifts were calculated using the DFT method. The chemical shifts were derived from the structures using the regression equation to convert shielding constants into chemical shifts. The calculated chemical shifts were then compared against NMR measured values. The experimental validation of the simulation data confirmed that the peptides exist in an ensemble of different substates rather than a single conformation. The ensemble R^2 values were always higher when compared to single conformations.

After it was established that enhanced sampling approaches can be used to probe conformational ensembles of AVP and OT, and capture their conformational dynamics, the same approach was applied to examine the conformational ensembles of two Urotensin peptides. Urotensin II and Urotensin Related Peptide differ in the total sequence size, but share the same ring residues. The URP peptide is characterised by only a single residue long N terminal, while UII has four residues in the N terminal tail.

The peptides have no determined crystal structures yet, and there is little other experimental data reported on their conformational characteristics. Therefore, the UII and URP cyclic peptides presented very good candidates to examine their conformational ensembles using enhanced sampling methods.

UII and URP showed more conformational flexibility than OT and AVP peptides. In total, eleven structural subtypes were identified for Urotensin peptides, with five of them highly populated for both peptides, while others were minor populated. In both peptides the same conformations were observed but with different populations. The Urotensin II most populated subtype was *Lasso*, while for Urotensin Related peptide it was *Open Omega*. However, both of the dominant conformational subtypes belong to the *open* conformational group, suggesting that UII and URP conformational dynamics is shifted towards more planar ring arrangements.

The biological data reported that UII and URP peptides bind to the same GPCR receptor but trigger different cascade reactions. In the context of our work, this could be rationalised by different dominant conformational subtypes, just like AVP and OT have the same dominant conformations found as their bound conformations, then UII and URP different dominant subtypes could also be connected with different receptor activation mechanisms.

Comparing the structural ensembles between all studied cyclic peptides, it can be seen that AVP and OT peptides prefer *folded* ring conformations stabilised with different hydrogen bonds. However, AVP is more conformational flexible than OT which seems to be only exchanging between two *folded* subtypes. On the other hand, UII and URP peptides are mostly found in the *open* ring conformations, each of them preferring different dominant conformational subtypes.

10.1 Cyclic peptide classification

All the peptides studied in this work share the same structural motif of a six-membered ring closed by a disulphide bridge between two cysteine residues. The structural classification based on adopted secondary motifs and populated hydrogen bonds found in the ring part of the peptide is a commonly used approach to report the structural diversity of the cyclic peptides.

Ring conformational classes. For the peptides studies in this work, two major ring structural classes were identified, named as *open* and *folded.* An *open*

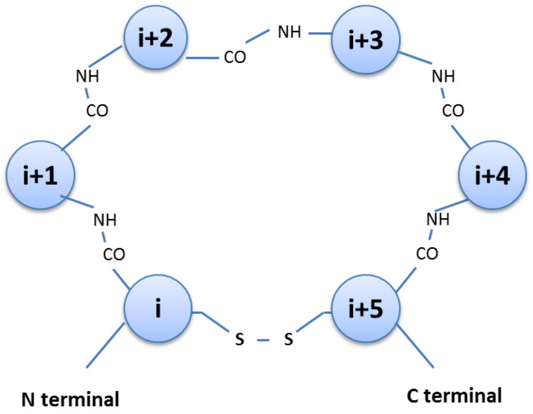


Figure 10.1: Conformational classification of peptide hormones with a 6-residue ring motif. Folded ring conformations show turns at residues i+2 and i+3, open (unfolded) ring conformations either at i+1,i+2 or i+3,i+4.

ring conformer was described as adopting β-turns centred at residues i+1, i+2 or i+3, i+4 with little or no populated intramolecular hydrogen bonds, while *folded* conformers share turns at residues i+2, i+3 stabilised with one or more intramolecular hydrogen bonds. Each of the two major structural classes were then described as adopting a number of subtypes.

The conformational subtypes of *open* and *folded* ring conformations were obtained using the torsion based clustering software Dash, later characterised by the adopted β-turn and hydrogen bonds.

For AVP and OT, there are four ring conformations identified:

- *Saddle, Open, Twisted Saddle, Clinched Open*

For UII and URP, there are in total eleven ring subtypes:

- *Omega I Open, Omega I hbond, Omega II, Lasso, Scoop, Hybrid (Folded IVb2), Sheet, Folded I, Folded II, Folded III, Inverted Folded*

Comparing the AVP, OT, UII and URP conformational ring subtypes, there are a few conformational similarities given in terms of the circular similarity between the ring torsion angles. The following *folded* ring substates were found to be similar between studied peptides:

- The AVP and OT *Saddle* ring conformation is 93 % circular similar to UII/URP *Folded I* ring conformation. The states were characterised with turn centres at residues i+2, i+3 with high population of $i^{+1}O-(i^{+4}H, i^{+5}H)$ hydrogen bonds

- The AVP/OT *Twisted Saddle* ring conformation is 91 and 90 % circular similar to UII/URP *Hybrid (Folded IVb2)* and *Sheet* ring conformation, respectively. The turn centres for this conformational state was also found at residues i+2, i+3, mostly stabilised with $i^{+1}O - i^{+4}H$ hydrogen bond

From the open ring conformational subtypes, the following subtypes were identified as similar:

- AVP/OT *Clinched Open* structure was 88 % similar to *Omega I Open* ring subtype. This ring conformation has turn centred at residues i+3, i+4 with no characteristic hydrogen bond

Overall, some of the identified ring conformations show very high structural similarity across all peptides suggesting that cyclic hexapeptides conformations can be in general described with the turns centred at particular residues irrespective of the ring sequence.

Tail conformations. The AVP, OT and UII tail conformations were described as *folded* or *extended*. If there was found a turn in the tail with its residues making a hydrogen bond with ring residues, then tail was described as *folded*. However, this tail conformation was not found to be significantly populated. In AVP and OT peptides the folded ring conformation was described to be populated 10 - 20 %. In UII peptide, the population of this tail conformation was higher, 20 - 30 % but the UII tail is also longer, consisting of 4 residues compared to AVP and OT tail made of 3 residues. Other tail conformations were described as *extended*, and they seem to be more favourable.

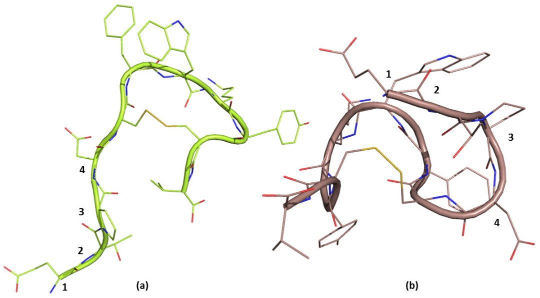


Figure 10.2: The example of the a) extended vs. b) folded UII tail orientation.

10.2 The efficiency of Replica Exchange methods

Finally, a comparison between the performance and efficiency of the advanced sampling methods used will be made. The conformational sampling of the Urotensin peptides was examined using both temperature Replica Exchange and Solute Tempering methods. The AVP conformational ensemble was explored using only temperature Replica Exchange, while the OT conformational dynamics was examined using only the Solute Tempering method.

The temperature Replica Exchange method is a more computationally demanding method than Solute Tempering. To obtain equilibrium sampling, a large number of replicas are required to cover a specified temperature range. In our case, to simulate a small peptide of only 9 amino acids in explicit solvent, 80 replicas were used to cover a temperature range 298 K to 550 K and to obtain 30 % acceptance probability between adjacent replicas. To study the conformational dynamics of the Urotensin peptides, a lower acceptance probability was obtained, 20 %, which required 64 replicas for the same temperature range as for the AVP peptide. The need for such a high number of replicas requires expensive parallel computing resources.

To tackle the high computational cost of the REMD method, a Solute Tempering method was applied that scales the solute - solute and solute - solvent interactions, enhancing sampling only of the "*hot*" parts of the system, the peptide in our case. As a consequence, a much smaller number of replicas is required to simulate the system across a range of temperatures, and to achieve an average acceptance probability of 0.2 - 0.3. In our work, the Urotensin peptides simulated with Solute Tempering were run with 12 replicas to achieve the acceptance probability between 20 - 30 %, while for the REMD run 64 replicas we used, which is a 5-fold decrease in the need for computational resources.

Despite being computational expensive, temperature Replica Exchange was extensively run for three peptides in this work, with four simulation repeats for AVP, and three for UII and URP peptides. The AVP simulation repeats were run for 300 ns per replica, URP for 400 ns, and UII for 500 ns per replica. In total, the simulations were run for 24 μs, 26 μs and 32 μs per simulation repeat for AVP, URP and OT respectively. The small cyclic peptides used in this work (8 - 11 residues, depending on the peptide), were shown to be conformationally complex even for the enhanced sampling method, requiring the need for the long simulation times. Moreover, the energy barrier required to see *cis/trans* isomerisation of the peptide bond involving proline residue, was not overcome in all simulation repeats for the temperature Replica Exchange method, suggesting that higher temperatures were needed to see steady *cis* populations across all simulation repeats.

To compare the efficiency of the enhanced sampling protocols, three repeats per method were run for each of the Urotensin peptides. The objective was to examine the potential advantages of the Solute Tempering method over REMD on the conformational sampling convergence times. The results obtained for both peptides show that a shorter simulation time is needed for Solute Tempering to get comparable substate populations compared to REMD simulations in a fraction of computational and real time cost, largely due to the fewer replicas needed. This result suggests that Solute Tempering simulations can be safely applied to study the conformational ensemble of intrinsically disordered peptides without affecting the sampling quality.

Chapter 11

Conclusions

In this thesis the conformational ensemble of the peptide hormones with a common structural motif of a six membered cyclic moiety closed by disulphide bridge was explored. A hypothesis that a combination of enhanced sampling methods together with the computationally calculated chemical shifts, compared against experimental chemical shifts, can be used to generate equilibrium ensemble was investigated.

The peptides studied in this thesis belong to the class of the recently discovered intrinsically disordered peptides. In Chapter 2, the review of the known structural and functional characteristics of the IDPs was given. The insight into their structural diversity is usually experimentally obtained using NMR, so in the remaining part of the Chapter 2, the theoretical basis of the measured NMR phenomena together with how the measured observables are translated into structural information was given.

Besides being experimentally determined, a detailed IDP structural characterisation can also be obtained using molecular dynamics methods. However, classical molecular dynamics simulation are known to suffer from sampling issue, so different advanced sampling methodologies have been applied to overcome the trapping issue. A review of several enhanced sampling methods was given in Chapter 3. The idea was to emphasise potential advantages and disadvantages of the most common enhanced sampling methods used in the literature to study the conformational ensembles of the peptides. In addition, the approach we used to investigate the peptide's inherent flexibility also consisted of calculating the chemical shifts using DFT, what was also explained in this chapter.

Finally, in the remaining chapters, the established methodological approach consisting of exploring peptide conformation using enhanced sampling methodology, and then validating the obtained ensemble against the experimental chemical shifts, was applied to all cyclic peptides. The results obtained show that the Replica Exchange methods are able to give us access to the conformational equilibria for all peptides. In addition, for the example of the Urotensin peptides whose conformational ensemble was extensively explored using two enhanced sampling methods and three repeats per method, it was shown that Solute Tempering is able to obtain a converged conformational ensemble faster due to the fewer replicas needed to run the method.

Next, to validate the extent to which the conformational ensemble is able to capture the IDP's conformational dynamics, the NMR measured ^{1}H chemical shifts were compared with the DFT calculated ^{1}H chemical shifts. This approach showed that the chemical shifts can be used to explore conformational equilibrium despite the fact that they are not very discriminating. A technique consisting of taking the chemical shifts as the sum of the population weighted chemical shifts was applied to all peptides studied in this thesis. The chemical shift weighted ensembles showed better agreement with experimental chemical shifts compared to calculation on the single representative structures, suggesting that simulations are yielding meaningful conformation ratios and populations.

The results obtained using the described methodology were further compared with the known experimental and biological data about the studied peptides. For the example of peptides with the known crystal structures, AVP and OT, it was shown that the converged AVP and OT conformational ensembles were able to obtain crystal structure conformations among the most populated conformational states. Both AVP and OT had their structures co-crystallised with different binding partners, and in both cases, the major populated conformational states in AVP and OT ensembles were their crystal structures. This suggests that when we do not know the crystal structure of the peptide, the bound conformation can be found from the conformational ensemble obtained in solution in the limit of converged simulation ensembles. Therefore, it can be assumed that the converged conformational ensemble of the Urotensin peptides also contains their bioactive conformation.

Furthermore, the tested methodology and results obtained are relevant in terms of development of the new therapeutics targeting GPCR receptors, as all the cyclic peptides studied in this work were found to exert their function through the activation of GPCR receptors. A conformational ensemble obtained using enhanced sampling methods contains a number of different conformations which could be relevant for drug design and serve as a template for docking calculations.

Appendix A

AVP and OT

		Tyr^2	Tyr^2	Phe^3	Phe^3	Gln^4	Gln^4	Asn^5	Asn^5	Cys^6	Cys^6
		ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ
Open		-108.82	136.51	53.23	3.74	-138.88	151.63	-76.01	129.26	-138.317	152.33
	stdev	33.20	19.44	22.40	30.34	26.74	18.98	18.32	25.64	25.05	27.11
Cl. Open		-83.51	-16.03	-120.81	153.35	-70.48	-23.40	-90.77	35.47	-90.40	142.42
	stdev	25.16	18.26	30.18	15.72	18.88	20.44	23.95	54.30	42.86	23.64
Saddle		-87.69	145.82	-60.58	-22.91	-86.80	-7.37	-116.03	-24.21	-123.14	141.36
	stdev	25.17	30.97	13.81	15.33	17.15	16.37	20.40	19.93	29.85	29.77
Tw. Saddle		-80.21	168.02	-52.82	125.84	56.74	7.40	-100.39	-18.35	-100.75	147.69
	stdev	35.89	14.37	15.24	15.27	10.34	25.16	25.57	25.45	31.37	21.01

Table A.1: The mean ring $\phi\psi$ torsion values of the representative AVP cluster states

		Tyr^2	Tyr^2	Ile^3	Ile^3	Gln^4	Gln^4	Asn^5	Asn^5	Cys^6	Cys^6
		ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ
Open		-80.93	130.98	42.63	-29.90	-102.34	143.52	-75.31	116.63	-95.25	140.09
	stdev	35.79	30.55	62.73	62.37	48.21	55.40	36.33	66.76	55.82	30.72
Cl. Open		-82.91	-24.12	-83.51	142.39	-64.52	-15.78	-75.48	13.90	-70.59	128.79
	stdev	30.42	72.51	65.78	54.28	33.24	65.17	65.26	50.17	59.83	36.27
Saddle		-75.39	146.73	-55.53	-31.72	-72.95	-15.61	-104.08	-18.16	-125.14	115.81
	stdev	25.51	23.60	26.57	32.61	35.82	40.47	45.23	33.17	46.51	53.86
Tw. Saddle		-78.81	155.37	-47.81	86.01	36.95	10.23	-100.91	-11.42	-93.24	137.78
	stdev	53.53	27.53	38.67	89.73	69.94	38.83	50.52	45.66	40.24	33.21

Table A.2: The mean ring $\phi\psi$ torsion values of the representative OXT cluster states

Appendix B

Urotensin Related Peptide

		Cys^2	Phe^3	Phe^3	Trp^4	Trp^4	Lys^5	Lys^5	Tyr^6	Tyr^6	Cys^7
		ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ
Omega I Open		143.51	-120.46	-17.75	-117.02	174.82	-56.73	-32.26	-138.30	141.36	-136.78
	stdev	13.93	29.39	, 23.69	32.52	15.52	21.34	26.82	29.70	25.89	23.54
Omega I hbond		141.00	-103.75	-4.59	-122.40	154.99	-58.17	-25.97	-85.27	1.55	-133.10
	stdev	34.21	42.33	48.11	40.80	21.91	22.31	22.01	23.66	39.83	30.90
Omega II		144.97	-95.03	, 0.91	-108.65	168.83	-58.25	153.67	56.76	33.84	-86.83
	stdev	25.92	24.37	32.66	39.13	23.31	26.95	24.21	19.94	23.51	40.26
Lasso		6.96	-62.75	-48.34	-131.64	-12.06	-100.99	159.94	-85.93	138.76	-129.36
	stdev	24.05	24.73	29.73	25.14	39.52	34.14	29.63	26.75	23.50	30.46
Folded I		146.18	-103.37	139.05	-57.06	-27.30	-68.90	-16.63	-130.26	-9.44	-122.22
	stdev	34.12	29.21	21.68	38.85	31.25	31.88	31.96	32.76	25.40	30.40
Inverted Folded		-1.65	-69.72	-28.85	-61.74	-24.68	-113.54	21.05	63.67	24.37	52.04
	stdev	13.18	20.12	12.40	17.18	13.50	20.57	13.97	11.54	34.77	34.67
Hybrid		135.05	-75.24	161.61	-43.67	135.11	64.06	21.59	-79.96	-23.45	-112.54
	stdev	20.46	27.16	15.28	, 27.05	29.32	28.12	17.96	15.88	17.87	37.89
Sheet		150.47	-110.99	-151.64	-73.84	97.87	57.40	-8.54	-139.35	136.65	-129.53
	stdev	18.30	34.75	34.03	27.82	49.50	9.06	39.46	21.60	34.64	21.67

Table B.1: The mean ring $\phi\psi$ torsion values of the representative URP cluster states

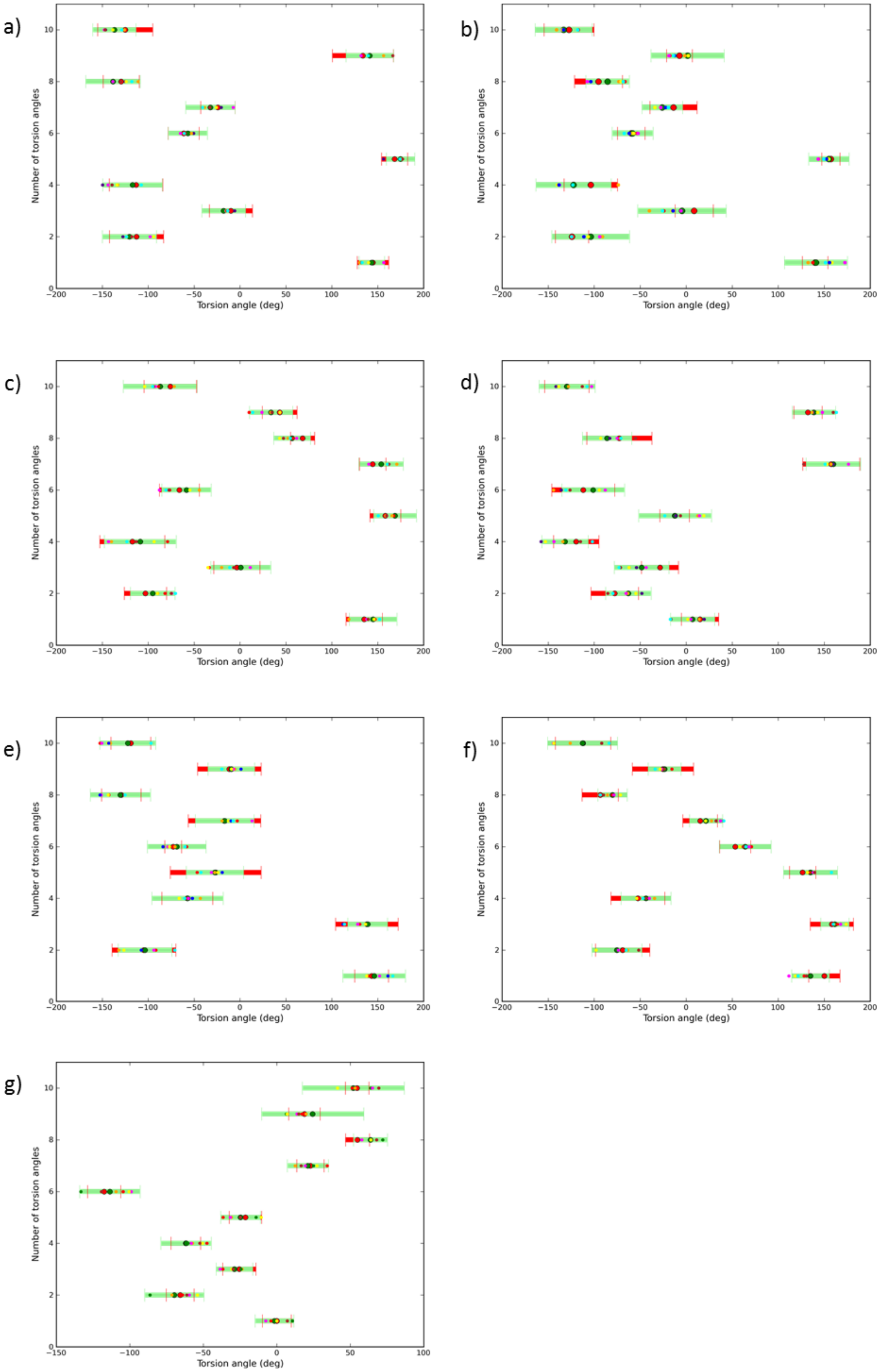


Figure B.1: The distribution of observed torsion angles for each URP conformation in each conformational state for a) Omega I Open, b) Omega I hbond, c) Omega II, d) Lasso, e) Folded I, f) Folded IVb2, g) Inverted Folded. The red bars are from MD simulations [182], and REMD torsion angle distributions are in green. The spots show the dihedral angles for the structures we selected.

Appendix C

Urotensin II

		Cys^5	Phe^6	Phe^6	Trp^7	Trp^7	Lys^8	Lys^8	Tyr^9	Tyr^9	Cys^{10}
		ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ	ψ	ϕ
Omega I Open		142.28	-125.19	-18.44	-105.46	170.27	-61.73	-20.31	-131.55	143.02	-125.92
	stdev	19.00	33.77	15.78	40.04	11.46	17.81	29.17	20.84	16.83	29.04
Omega I hbond		138.51	-111.67	-5.01	-96.10	147.85	-59.59	-20.38	-91.18	-0.38	-128.83
	stdev	26.90	26.82	23.17	35.65	39.39	22.03	23.22	26.41	20.55	40.75
Omega II		141.77	-70.84	-15.82	-108.94	152.64	-73.75	151.36	55.94	41.79	-82.98
	stdev	20.33	17.73	23.90	27.51	34.45	15.36	11.07	8.01	25.21	14.26
Lasso		8.49	-79.75	-26.89	-121.15	-19.83	-105.19	155.19	-74.29	130.33	-131.85
	stdev	31.28	23.33	14.47	19.88	17.53	29.66	13.24	12.47	16.71	18.32
Scoop		169.31	-58.71	-47.17	-82.21	21.33	62.23	-49.85	-133.37	132.27	-133.14
	stdev	11.83	8.18	9.08	9.96	13.63	7.36	22.12	14.95	12.40	10.42
Circle		25.415	-136.77	-24.90	-134.18	-36.63	-128.81	-39.04	-142.38	147.64	-155.05
	stdev	16.03	25.74	16.22	22.07	24.53	, 21.24	35.58	40.52	15.57	55.94
Folded I		117.38	-98.82	140.63	-57.06	-24.30	-76.92	-20.61	-130.26	-5.25	-135.67
	stdev	56.91	28.93	20.53	8.85	13.25	18.88	13.97	12.75	30.46	36.03
Folded II		-12.80	, -62.66	134.05	57.17	4.20	-103.66	-45.60	-149.38	-15.72	-104.54
	stdev	15.74	20.00	15.37	15.24	24.95	27.36	20.24	11.21	25.65	25.08
Folded III		15.55	57.11	31.40	51.38	9.73	-127.15	-33.55	-104.45	15.87	-109.95
	stdev	15.57	13.21	13.72	13.29	19.99	20.90	15.53	17.90	15.65	33.27
Folded IVb2		144.09	-62.59	156.80	-49.74	125.25	55.55	12.02	-80.61	-25.45	-114.40
	stdev	36.35	13.53	13.06	15.41	11.44	15.98	17.25	20.10	18.93	28.80
Inverted Folded		-1.65	-69.72	-28.85	-61.74	-24.68	-113.54	21.05	63.67	24.37	32.04
	stdev	13.18	20.12	12.40	17.18	13.50	20.57	13.97	11.54	34.77	41.67

Table C.1: The mean ring $\phi\psi$ torsion values of the representative UII cluster states

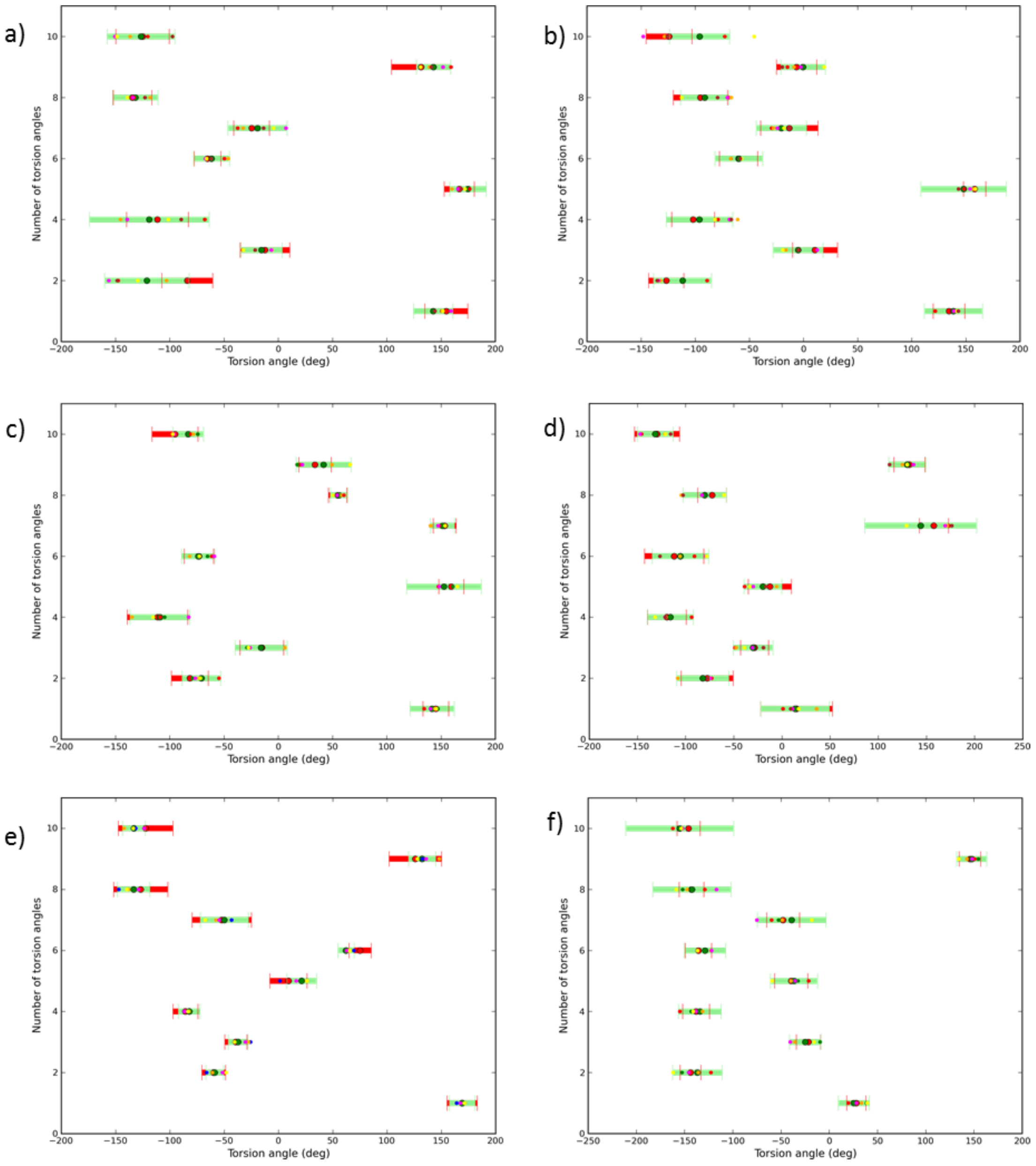


Figure C.1: The distribution of observed torsion angles for each UII conformation in each conformational state for a) Omega I Open, b) Omega I hbond, c) Omega II, d) Lasso, e) Scoop, f) Circle. The red bars are from MD simulations [182], and REMD torsion angle distributions are in green. The spots show the dihedral angles for the structures we selected.

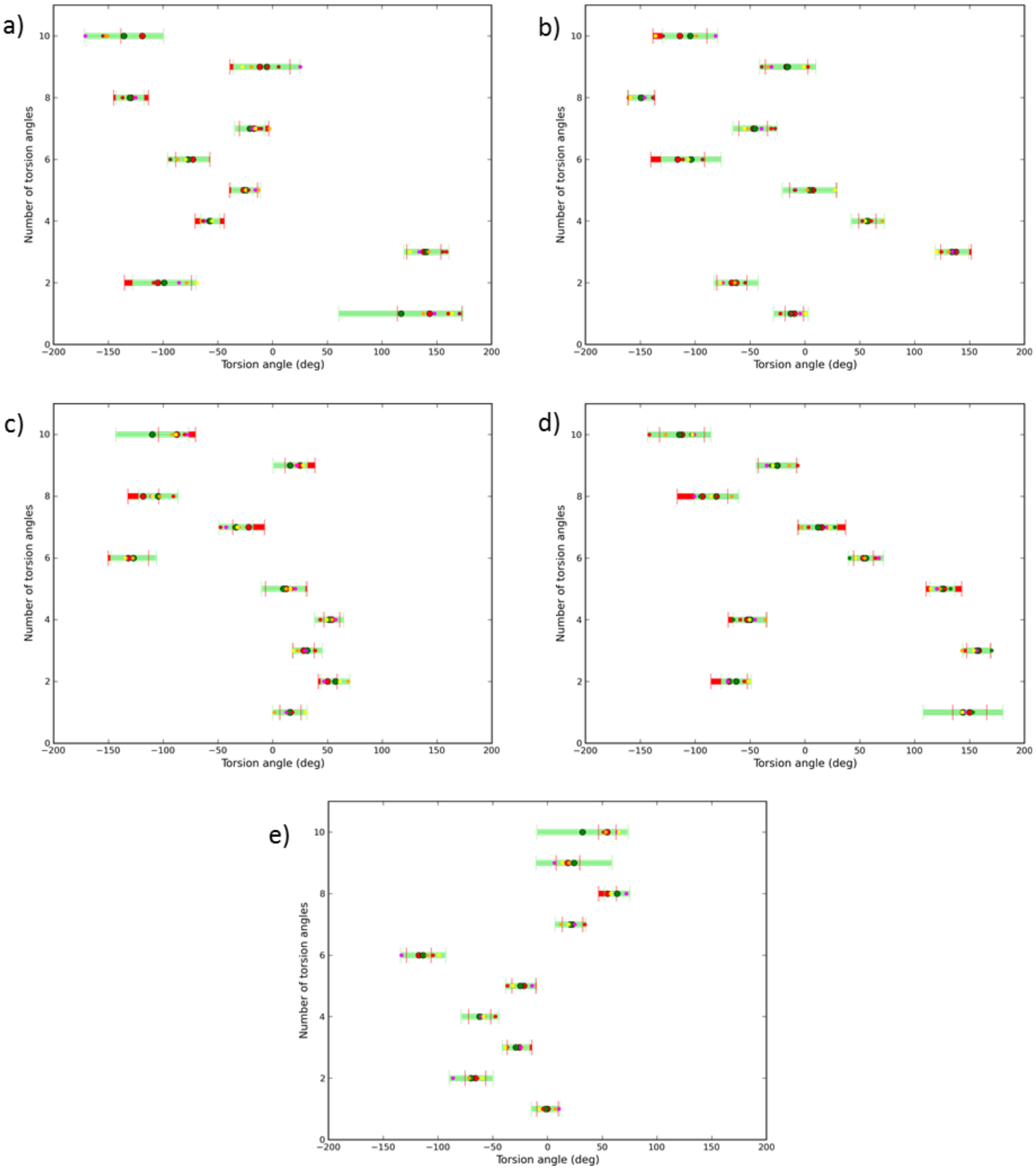


Figure C.2: The distribution of observed torsion angles for each UII conformation in each conformational state for a) Folded I, b) Folded II, c) Folded III, d) Folded IVb2, e) Inverted Folded. The red bars are from MD simulations [182], and REMD torsion angle distributions are in green. The spots show the dihedral angles for the structures we selected.

Bibliography

[1] C. M. Nelson DL, Lehninger AL, *Lehninger principles of biochemistry*, Macmillan, USA, **2008**.

[2] P. Andrews, *Biopolymers* **1971**, *10*, 2253–2267.

[3] G. Scherer, M. L. Kramer, M. Schutkowski, U. Reimer, G. Fischer, *Journal of the American Chemical Society* **1998**, *120*, 5568–5574.

[4] C Ramakrishnan, P. K. Paul, K Ramnarayan, *Journal of Biosciences* **1985**, *8*, 239–251.

[5] K. Wuthrich, C. Grathwohl, *FEBS letters* **1974**, *43*, 337–340.

[6] G. N. Ramachandran, C. Ramakrishnan, V Sasisekharan, *Journal of molecular biology* **1963**, *7*, 95–99.

[7] J. M. Berg, J. L. Tymoczko, L. Stryer, *Biochemistry. 5th*, WH Freeman, USA, **2002**.

[8] D. Voet, J. G. Voet, *Biochemistry*, John Wiley & Sons, USA, **2004**.

[9] J. S. Richardson, *Advances in protein chemistry* **1981**, *34*, 167–339.

[10] L. Pauling, R. B. Corey, H. R. Branson, *Proceedings of the National Academy of Sciences* **1951**, *37*, 205–211.

[11] J. S. Richardson, *Nature* **1977**, *268*, 495–500.

[12] C. Venkatachalam, *Biopolymers* **1968**, *6*, 1425–1436.

[13] P. N. Lewis, F. A. Momany, H. A. Scheraga, *Biochimica et Biophysica Acta (BBA)-Protein Structure* **1973**, *303*, 211–229.

[14] G. D. Rose, L. M. Glerasch, J. A. Smith, *Advances in protein chemistry* **1985**, *37*, 1–109.

[15] β-turn types, **2017**, `http://www.cryst.bbk.ac.uk/PPS2/course/section8/ss-960531-16.html` (visited on 10/22/2010).

[16] B. Matthews, *Macromolecules* **1972**, *5*, 818–819.

[17] E. J. Milner-White, *Journal of molecular biology* **1990**, *216*, 385–397.

[18] E. J. Milner-White, B. M. Ross, R. Ismail, K. Belhadj-Mostefa, R. Poet, *Journal of molecular biology* **1988**, *204*, 777–782.

[19] R. E. Hubbard, M. Kamran Haider, *Hydrogen bonds in proteins: role and strength*, Wiley Online Library, USA, **2010**.

[20] V. J. Hruby, *Life sciences* **1982**, *31*, 189–199.

[21] A. Tapeinou, M.-T. Matsoukas, C. Simal, T. Tselios, *Peptide Science* **2015**, *104*, 453–461.

[22] K. D. Kopple, *Journal of Pharmaceutical Sciences* **1972**, *61*, 1345–1356.

[23] A. Piserchio, G. D. Salinas, T. Li, J. Marshall, M. R. Spaller, D. F. Mierke, *Chemistry & biology* **2004**, *11*, 469–473.

[24] K. Shibata, T. Suzawa, S. Soga, T. Mizukami, K. Yamada, N. Hanai, M. Yamasaki, *Bioorganic & medicinal chemistry letters* **2003**, *13*, 2583–2586.

[25] T. A. Cardote, A. Ciulli, *ChemMedChem* **2016**, *11*, 787–794.

[26] C. M. Stegmann, R. Luhrmann, M. C. Wahl, *PloS one* **2010**, *5*, e10013.

[27] D. P. Fairlie, G. Abbenante, D. R. March, *Current medicinal chemistry* **1995**, *2*, 654–686.

[28] M. L. Korsinczky, H. J. Schirra, K. J. Rosengren, J. West, B. A. Condie, L. Otvos, M. A. Anderson, D. J. Craik, *Journal of molecular biology* **2001**, *311*, 579–591.

[29] M. Trabi, H. J. Schirra, D. J. Craik, *Biochemistry* **2001**, *40*, 4211–4221.

[30] D. J. Craik, N. L. Daly, T. Bond, C. Waine, *Journal of molecular biology* **1999**, *294*, 1327–1336.

[31] M. B. Martins, I. Carvalho, *Tetrahedron* **2007**, *63*, 9923–9932.

[32] M. T. Oakley, E. Oheix, A. F. Peacock, R. L. Johnston, *The journal of physical chemistry B* **2013**, *117*, 8122–8134.

[33] H. Kessler, *Angewandte Chemie International Edition* **1982**, *21*, 512–523.

[34] A. Jabs, M. S. Weiss, R. Hilgenfeld, *Journal of molecular biology* **1999**, *286*, 291–304.

[35] Y. Che, G. R. Marshall, *Journal of medicinal chemistry* **2006**, *49*, 111–124.

[36] M. T. Oakley, R. L. Johnston, *Journal of chemical theory and computation* **2012**, *9*, 650–657.

[37] C. M. Deber, V. Madison, E. R. Blout, *Accounts of Chemical Research* **1976**, *9*, 106–113.

[38] T. Rezai, J. E. Bock, M. V. Zhou, C. Kalyanaraman, R. S. Lokey, M. P. Jacobson, *Journal of the American Chemical Society* **2006**, *128*, 14073–14080.

[39] S. M. McHugh, J. R. Rogers, H. Yu, Y.-S. Lin, *Journal of chemical theory and computation* **2016**, *12*, 2480–2488.

[40] S. M. McHugh, J. R. Rogers, S. A. Solomon, H. Yu, Y.-S. Lin, *Current opinion in chemical biology* **2016**, *34*, 95–102.

[41] A. K. Yudin, *Chemical science* **2015**, *6*, 30–49.

[42] Hormones, **2018**, `http://www2.highlands.edu/academics/divisions/scipe/biology/faculty/harnden/2122/notes/endo.htm` (visited on 03/18/2018).

[43] E Fischer, *Berichte der Deutschen Chemischen Gesellschaft* **1894**, *27*, 2985–2993.

[44] D. Koshland, *Proceedings of the National Academy of Sciences* **1958**, *44*, 98–104.

[45] B. K. Kobilka, *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2007**, *1768*, 794–807.

[46] R. Schwyzer, *Journal of Molecular Recognition* **1995**, *8*, 3–8.

[47] S. Moro, C. Hoffmann, K. A. Jacobson, *Biochemistry* **1999**, *38*, 3498–3507.

[48] L. M. Luttrell, S. Maudsley, L. M. Bohn, *Molecular pharmacology* **2015**, *88*, 579–588.

[49] T. A. Kohout, S. L. Nicholas, S. J. Perry, G. Reinhart, S. Junger, R. S. Struthers, *Journal of Biological Chemistry* **2004**, *279*, 23214–23222.

[50] D. Chatenet, T. T. M. Nguyen, M. Letourneau, A. Fournier, *Frontiers in endocrinology* **2013**, *3*, 174.

[51] D. Brancaccio, F. Merlino, A. Limatola, A. M. Yousif, I. Gomez-Monterrey, P. Campiglia, E. Novellino, P. Grieco, A. Carotenuto, *Journal of Peptide Science* **2015**, *21*, 392–399.

[52] F. Shahidi, Y. Zhong, *Journal of AOAC International* **2008**, *91*, 914–931.

[53] T. Breuder, C. S. Hemenway, N. R. Movva, M. E. Cardenas, J. Heitman, *Proceedings of the National Academy of Sciences* **1994**, *91*, 5372–5376.

[54] E. K. Schmitt, M. Riwanto, V. Sambandamurthy, S. Roggo, C. Miault, C. Zwingelstein, P. Krastel, C. Noble, D. Beer, S. P. Rao, et al., *Angewandte Chemie International Edition* **2011**, *50*, 5889–5891.

[55] A. M. Blanks, S. Thornton, *BJOG: An International Journal of Obstetrics & Gynaecology* **2003**, *110*, 46–51.

[56] R. I. Lehrer, A. M. Cole, M. E. Selsted, *Journal of Biological Chemistry* **2012**, *287*, 27014–27019.

[57] I.-J. Ryoo, H.-R. Park, S.-J. Choo, J.-H. Hwang, Y.-M. Park, K.-H. Bae, K. Shin-Ya, I.-D. Yoo, *Biological and Pharmaceutical Bulletin* **2006**, *29*, 817–820.

[58] D. T. Krieger, *Science* **1983**, *222*, 975–985.

[59] L. Nevola, E. Giralt, *Chemical Communications* **2015**, *51*, 3302–3315.

[60] A. Zorzi, K. Deyle, C. Heinis, *Current Opinion in Chemical Biology* **2017**, *38*, 24–29.

[61] J. Renukuntla, A. D. Vadlapudi, A. Patel, S. H. Boddu, A. K. Mitra, *International journal of pharmaceutics* **2013**, *447*, 75–93.

[62] O. Ovadia, S. Greenberg, J. Chatterjee, B. Laufer, F. Opperer, H. Kessler, C. Gilon, A. Hoffman, *Molecular pharmaceutics* **2011**, *8*, 479–487.

[63] D. J. Craik, D. P. Fairlie, S. Liras, D. Price, *Chemical biology & drug design* **2013**, *81*, 136–147.

[64] Y. Fang, T. Kenakin, C. Liu, *Frontiers in pharmacology* **2015**, *6*.

[65] P. Vlieghe, V. Lisowski, J. Martinez, M. Khrestchatisky, *Drug discovery today* **2010**, *15*, 40–56.

[66] V. N. Uversky, *Chemical Reviews* **2014**, *114*, 6557—6560.

[67] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, et al., *Journal of Molecular Graphics and Modelling* **2001**, *19*, 26–59.

[68] V. N. Uversky, J. R. Gillespie, A. L. Fink, *Proteins: structure function and bioinformatics* **2000**, *41*, 415–427.

[69] V. N. Uversky, A. K. Dunker, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2010**, *1804*, 1231–1264.

[70] L. Mollica, L. M. Bessa, X. Hanoulle, M. R. Jensen, M. Blackledge, R. Schneider, *Frontiers in molecular biosciences* **2016**, *3*, 52–70.

[71] H. J. Dyson, P. E. Wright, *Current opinion in structural biology* **2002**, *12*, 54–60.

[72] D. D. Boehr, R. Nussinov, P. E. Wright, *Nature chemical biology* **2009**, *5*, 789–796.

[73] M. Kjaergaard, K. Teilum, F. M. Poulsen, *Proceedings of the National Academy of Sciences* **2010**, *107*, 12535–12540.

[74] S. Fermani, X. Trivelli, F. Sparla, A. Thumiger, M. Calvaresi, L. Marri, G. Falini, F. Zerbetto, P. Trost, *Journal of Biological Chemistry* **2012**, *287*, 21372–21383.

[75] M. Arai, K. Sugase, H. J. Dyson, P. E. Wright, *Proceedings of the National Academy of Sciences* **2015**, *112*, 9614–9619.

[76] B. A. Shoemaker, J. J. Portman, P. G. Wolynes, *Proceedings of the National Academy of Sciences* **2000**, *97*, 8868–8873.

[77] H. J. Dyson, P. E. Wright, *Nature reviews. Molecular cell biology* **2005**, *6*, 197–208.

[78] N. G. Sgourakis, Y. Yan, S. A. McCallum, C. Wang, A. E. Garcia, *Journal of molecular biology* **2007**, *368*, 1448–1457.

[79] H. J. Dyson, P. E. Wright, *Methods in enzymology* **2000**, *339*, 258–270.

[80] D. Kruschel, B. Zagrovic, *Molecular Biosystems* **2009**, *5*, 1606–1616.

[81] J. B. Lambert, E. P. Mazzola, *Nuclear magnetic resonance spectroscopy: an introduction to principles, applications, and experimental methods*, Pearson education, USA, **2004**.

[82] P. Atkins, J. De Paula, *Elements of physical chemistry*, Oxford University Press, USA, **2013**.

[83] D. S. Wishart, B. D. Sykes, F. M. Richards, *Biochemistry* **1992**, *31*, 1647–1651.

[84] M. Karplus, *The Journal of chemical physics* **1959**, *30*, 11–15.

[85] M. J. Minch, *Concepts in Magnetic Resonance Part A* **1994**, *6*, 41–56.

[86] A. W. Overhauser, *Physical Review* **1953**, *92*, 411.

[87] N. Tjandra, A. Bax, *Science* **1997**, *278*, 1111–1114.

[88] E. Brunner, *Concepts in Magnetic Resonance Part A* **2001**, *13*, 238–259.

[89] K. Chen, N. Tjandra, *eMagRes* **2011**, *4*, 47–67.

[90] A. G. Palmer III, *Chemical reviews* **2004**, *104*, 3623–3640.

[91] H. Kessler, *Angewandte Chemie International Edition* **1970**, *9*, 219–235.

[92] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, D. E. Shaw, *Annual review of biophysics* **2012**, *41*, 429–452.

[93] K. Henzler-Wildman, D. Kern, *Nature* **2007**, *450*, 964–972.

[94] R. A. Friesner, *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 6648–6653.

[95] A. Einstein, *Annalen der physik* **1905**, *322*, 132–148.

[96] D. J. Griffiths, *Introduction to quantum mechanics*, Cambridge University Press, UK, **2016**.

[97] V. Fock, *Zeitschrift fur Physik A Hadrons and Nuclei* **1930**, *61*, 126–148.

[98] D. R. Hartree, *Mathematical Proceedings of the Cambridge Philosophical Society* **1928**, *24*, 89–110.

[99] P.-O. Lowdin, *Physical review* **1955**, *97*, 1509–1520.

[100] P. Hohenberg, W. Kohn, *Physical review* **1964**, *136*, B864.

[101] W. Kohn, L. J. Sham, *Physical review* **1965**, *140*, A1133.

[102] A. D. Becke, *The Journal of chemical physics* **1993**, *98*, 1372–1377.

[103] P. Stephens, F. Devlin, C. Chabalowski, M. J. Frisch, *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

[104] J. C. Slater, *Physical Review* **1930**, *36*, 57–64.

[105] S. F. Boys, *Proceedings of the Royal Society of London A: Mathematical Physical and Engineering Sciences* **1950**, *200*, 542–554.

[106] B. Nagy, F. Jensen, *Reviews in Computational Chemistry* **2017**, 93–149.

[107] R. Ditchfield, W. J. Hehre, J. A. Pople, *The Journal of Chemical Physics* **1971**, *54*, 724–728.

[108] B. Han, Y. Liu, S. W. Ginzinger, D. S. Wishart, *Journal of biomolecular NMR* **2011**, *50*, 43–57.

[109] Y. Shen, A. Bax, *Journal of biomolecular NMR* **2010**, *48*, 13–22.

[110] K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, M. Vendruscolo, *Journal of the American Chemical Society* **2009**, *131*, 13894–13895.

[111] R. Jain, T. Bally, P. R. Rablen, *The Journal of organic chemistry* **2009**, *74*, 4017–4023.

[112] M. W. Lodewyk, M. R. Siebert, D. J. Tantillo, *Chemical reviews* **2011**, *112*, 1839–1862.

[113] W. Kutzelnigg, U. Fleischer, C. van Wullen, *Encyclopedia of Nuclear Magnetic Resonance* **1996**, *7*, 4284–4291.

[114] K. Wolinski, J. F. Hinton, P. Pulay, *Journal of the American Chemical Society* **1990**, *112*, 8251–8260.

[115] J. C. Facelli, *Progress in nuclear magnetic resonance spectroscopy* **2011**, *58*, 176–201.

[116] M. Frisch, G. Trucks, H. B. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G Scalmani, V Barone, B Mennucci, G. Petersson, et al., Gaussian 09, revision D. 01, **2009**.

[117] N. van Eikema Hommes, T. Clark, *Journal of molecular modeling* **2005**, *11*, 175–185.

[118] J. Tomasi, B. Mennucci, R. Cammi, *Chemical Reviews* **2005**, *105*, 2999–3094.

[119] L. Verlet, *Physical review* **1967**, *159*, 98–103.

[120] R. W. Hockney, *Methods of Computational Physics* **1970**, *9*, 136–2011.

[121] J.-P. Ryckaert, G. Ciccotti, H. J. Berendsen, *Journal of Computational Physics* **1977**, *23*, 327–341.

[122] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc* **1996**, *118*, 11225–11236.

[123] C. Oostenbrink, A. Villa, A. E. Mark, W. F. Van Gunsteren, *Journal of computational chemistry* **2004**, *25*, 1656–1676.

[124] B. R. Brooks, C. L. Brooks, A. D. MacKerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al., *Journal of computational chemistry* **2009**, *30*, 1545–1614.

[125] D. A. Case, V Babin, J. Berryman, R. Betz, Q Cai, D. Cerutti, T. Cheatham Iii, T. Darden, R. Duke, H Gohlke, et al., **2014**.

[126] T. Darden, D. York, L. Pedersen, *The Journal of chemical physics* **1993**, *98*, 10089–10092.

[127] H. C. Andersen, *The Journal of chemical physics* **1980**, *72*, 2384–2393.

[128] W. G. Hoover, *Physical review A* **1985**, *31*, 1695–1697.

[129] H. Berendsen, J. Postma, W. van Gunsteren, A Dinola, J. Haak, *J. Chem. Phys*, *81*, 571–572.

[130] H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola, J. Haak, *The Journal of chemical physics* **1984**, *81*, 3684–3690.

[131] S. Piana, K. Lindorff-Larsen, D. E. Shaw, *Proceedings of the National Academy of Sciences* **2012**, *109*, 17845–17850.

[132] R. C. Bernardi, M. C. Melo, K. Schulten, *Biochimica et Biophysica Acta (BBA)-General Subjects* **2015**, *1850*, 872–877.

[133] Y. Miao, J. A. McCammon, *Molecular simulation* **2016**, *42*, 1046–1055.

[134] K. Ostermeir, M. Zacharias, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **2013**, *1834*, 847–853.

[135] G. M. Torrie, J. P. Valleau, *Journal of Computational Physics* **1977**, *23*, 187–199.

[136] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, *Journal of computational chemistry* **1992**, *13*, 1011–1021.

[137] A. Laio, M. Parrinello, *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

[138] C. Abrams, G. Bussi, *Entropy* **2013**, *16*, 163–199.

[139] A. Laio, F. L. Gervasio, *Reports on Progress in Physics* **2008**, *71*, 126601.

[140] A. Barducci, G. Bussi, M. Parrinello, *Physical review letters* **2008**, *100*, 020603.

[141] G. Bussi, F. L. Gervasio, A. Laio, M. Parrinello, *Journal of the American Chemical Society* **2006**, *128*, 13435–13441.

[142] D. Hamelberg, J. Mongan, J. A. McCammon, *The Journal of chemical physics* **2004**, *120*, 11919–11929.

[143] P. R. Markwick, J. A. McCammon, *Physical Chemistry Chemical Physics* **2011**, *13*, 20053–20065.

[144] L. C. Pierce, R. Salomon-Ferrer, C. Augusto F. de Oliveira, J. A. McCammon, R. C. Walker, *Journal of Chemical Theory and Computation* **2012**, *8*, 2997–3002.

[145] K. Kappel, Y. Miao, J. A. McCammon, *Quarterly reviews of biophysics* **2015**, *48*, 479–487.

[146] S. Mukherjee, R. K. Kar, R. P. R. Nanga, K. H. Mroue, A. Ramamoorthy, A. Bhunia, *Physical Chemistry Chemical Physics* **2017**, *19*, 19289–19299.

[147] Y. Sugita, Y. Okamoto, *Chemical physics letters* **1999**, *314*, 141–151.

[148] P. Liu, B. Kim, R. A. Friesner, B. Berne, *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 13749–13754.

[149] W. Zhang, C. Wu, Y. Duan, *The Journal of chemical physics* **2005**, *123*, 154105.

[150] N. Rathore, M. Chopra, J. J. de Pablo, *The Journal of chemical physics* **2005**, *122*, 024111.

[151] H. Fukunishi, O. Watanabe, S. Takada, *The Journal of chemical physics* **2002**, *116*, 9058–9067.

[152] L. Wang, R. A. Friesner, B. Berne, *The Journal of Physical Chemistry B* **2011**, *115*, 9431–9438.

[153] A. K. Smith, C. Lockhart, D. K. Klimov, *Journal of Chemical Theory and Computation* **2016**, *12*, 5201–5214.

[154] A. H. Brown, P. M. Rodger, J. S. Evans, T. R. Walsh, *Biomacromolecules* **2014**, *15*, 4467–4479.

[155] G. Bussi, *Molecular Physics* **2014**, *112*, 379–384.

[156] D. W. Salt, B. D. Hudson, L. Banting, M. J. Ellis, M. G. Ford, *Journal of medicinal chemistry* **2005**, *48*, 3214–3220.

[157] B. Efron, R. J. Tibshirani, *An introduction to the bootstrap*, CRC press, USA, **1994**.

[158] V. du Vigneaud, D. T. Gish, P. G. Katsoyannis, *Journal of the American Chemical Society* **1954**, *76*, 4751–4752.

[159] B. S. Ibrahim, V. Pattabhi, *Journal of molecular biology* **2005**, *348*, 1191–1198.

[160] S. Nielsen, C.-L. Chou, D. Marples, E. I. Christensen, B. K. Kishore, M. A. Knepper, *Proceedings of the National Academy of Sciences* **1995**, *92*, 1013–1017.

[161] ADH, **2017**, `https://www.myvmc.com/diseases/syndrome-of-inappropriate-antidiuretic-hormone-secretion-siadh/` (visited on 10/23/2010).

[162] M. Miller, T. Dalakos, A. M. Moses, H. Fellerman, D. Streeten, *Annals of Internal Medicine* **1970**, *73*, 721–9.

[163] A. G. Robinson, *New England Journal of Medicine* **1976**, *294*, 507–511.

[164] Q. Pittman, B Bagdan, *Progress in brain research* **1992**, *91*, 69–74.

[165] D. W. Wacker, M. Ludwig, *Hormones and behavior* **2012**, *61*, 259–265.

[166] C. K. Wu, B. Hu, J. P. Rose, Z.-J. Liu, T. L. Nguyen, C. Zheng, E. Breslow, B.-C. Wang, *Protein Science* **2001**, *10*, 1869–1880.

[167] E. Sikorska, S. Rodziewicz-MotowidŁo, *Journal of Peptide Science* **2008**, *14*, 76–84.

[168] J. M. Scmidth, O. Ohlenschlager, H. Ruterjans, Z. Grzonka, E. Kojro, I. Pavo, F. Fahrenholz, *The FEBS Journal* **1991**, *201*, 355–371.

[169] S. Rodziewicz-Motowidlo, E. Sikorska, M. Oleszczuk, C. Czaplewski, *Journal of Peptide Science* **2008**, *14*, 85–96.

[170] C Barberis, B Mouillac, T Durroux, *Journal of Endocrinology* **1998**, *156*, 223–229.

[171] E. A. Lubecka, E. Sikorska, D. Sobolewski, A. Prahl, J. Slaninova, J. Ciarkowski, *European Biophysics Journal* **2015**, *44*, 727–743.

[172] A. Liwo, A. Tempczyk, S. Oldziej, M. D. Shenderovich, V. J. Hruby, S. Talluri, J. Ciarkowski, F. Kasprzykowski, L. Lankiewicz, Z. Grzonka, *Biopolymers* **1996**, *38*, 157–175.

[173] E. Haensele, L. Banting, D. C. Whitley, T. Clark, *Journal of molecular modeling* **2014**, *20*, 2485.

[174] E. Yedvabny, P. S. Nerenberg, C. So, T. Head-Gordon, *The Journal of Physical Chemistry B* **2014**, *119*, 896–905.

[175] A. Okur, D. R. Roe, G. Cui, V. Hornak, C. Simmerling, *Journal of chemical theory and computation* **2007**, *3*, 557–568.

[176] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, R. J. Woods, *Journal of computational chemistry* **2005**, *26*, 1668–1688.

[177] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *The Journal of chemical physics* **1983**, *79*, 926–935.

[178] A. Patriksson, D. van der Spoel, *Physical Chemistry Chemical Physics* **2008**, *10*, 2073–2077.

[179] C. K. Larive, L. Guerra, D. L. Rabenstein, *Journal of the American Chemical Society* **1992**, *114*, 7331–7337.

[180] E. Haensele, N. Saleh, C. M. Read, L. Banting, D. C. Whitley, T. Clark, *Journal of chemical information and modeling* **2016**, *56*, 1798–1807.

[181] E. Benassi, *Journal of computational chemistry* **2017**, *38*, 87–92.

[182] E. Haensele, PhD thesis, University of Portsmouth, **2017**.

[183] V. d. Vigneaud, C. Ressler, C. J. M. Swan, C. W. Roberts, P. G. Katsoyannis, S. Gordon, *Journal of the American Chemical Society* **1953**, *75*, 4879–4880.

[184] D. Jarvis, V. Du Vigneaud, *Science* **1964**, *143*, 545–548.

[185] B. M. Ferrier, D. Jarvis, V. Du Vigneaud, *Journal of Biological Chemistry* **1965**, *240*, 4264–4266.

[186] B. W. Low, C. C. Chen, *Science* **1966**, *151*, 1552–1553.

[187] S. Wood, *Science* **1986**, *232*, 633–637.

[188] J Husain, T. Blundell, S Cooper, J. Pitts, I. Tickle, S. Wood, V. Hruby, A Buku, A. Fischman, H. Wyssbrod, et al., *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **1990**, *327*, 625–654.

[189] J. P. Rose, C.-K. Wu, C.-D. Hsiao, E. Breslow, B.-C. Wang, *Nature Structural & Molecular Biology* **1996**, *3*, 163–169.

[190] A. Ohno, N. Kawasaki, K. Fukuhara, H. Okuda, T. Yamaguchi, *Magnetic Resonance in Chemistry* **2010**, *48*, 168–172.

[191] J. Koehbach, M. O'Brien, M. Muttenthaler, M. Miazzo, M. Akcan, A. G. Elliott, N. L. Daly, P. J. Harvey, S. Arrowsmith, S. Gunasekera, et al., *Proceedings of the National Academy of Sciences* **2013**, *110*, 21183–21188.

[192] M Budesinsky, U Ragnarsson, L Lankiewicz, L Grehn, J Slaninova, J Hlavacek, *Amino acids* **2005**, *29*, 151–160.

[193] R Bhaskaran, L.-C. Chuang, C. Yu, *Biopolymers* **1992**, *32*, 1599–1608.

[194] T. Kato, S. Endo, T. Fujiwara, K. Nagayama, *Journal of biomolecular NMR* **1993**, *3*, 653–673.

[195] M. Birnbaumer, *Trends in Endocrinology & Metabolism* **2000**, *11*, 406–410.

[196] T. Kimura, O. Tanizawa, et al., *Nature* **1992**, *356*, 526–529.

[197] R. Postina, E. Kojro, F. Fahrenholz, *Journal of Biological Chemistry* **1996**, *271*, 31593–31601.

[198] N. Saleh, G. Saladino, F. L. Gervasio, E. Haensele, L. Banting, D. C. Whitley, J. Sopkova-de Oliveira Santos, R. Bureau, T. Clark, *Angewandte Chemie* **2016**, *128*, 8140–8144.

[199] M. J. Slusarz, A. Gieldon, R. Slusarz, J. Ciarkowski, *Journal of Peptide Science* **2006**, *12*, 180–189.

[200] M Manning, A Misicka, A Olma, K Bankowski, S Stoev, B Chini, T Durroux, B Mouillac, M Corbani, G Guillon, *Journal of neuroendocrinology* **2012**, *24*, 609–628.

[201] M. J. Slusarz, R. Slusarz, J. Ciarkowski, *Journal of Peptide Science* **2006**, *12*, 171–179.

[202] C.-Y. Ku, A. Qian, Y. Wen, K. Anwer, B. M. Sanborn, *Endocrinology* **1995**, *136*, 1509–1515.

[203] J. M. Krieger, G. Fusco, M. Lewitzky, P. C. Simister, J. Marchant, C. Camilloni, S. M. Feller, A. De Simone, *Biophysical journal* **2014**, *106*, 1771–1779.

[204] H. Bern, K Lederis, *The Journal of endocrinology* **1969**, *45*, 341–349.

[205] M. Mori, T. Sugo, M. Abe, Y. Shimomura, M. Kurihara, C. Kitada, K. Kikuchi, Y. Shintani, T. Kurokawa, H. Onda, et al., *Biochemical and biophysical research communications* **1999**, *265*, 123–129.

[206] H.-P. Nothacker, Z. Wang, A. M. McNeill, Y. Saito, S. Merten, B. O'Dowd, S. P. Duckles, O. Civelli, *Nature Cell Biology* **1999**, *1*, 383–385.

[207] Q. Liu, S.-S. Pong, Z. Zeng, Q. Zhang, A. D. Howard, D. L. Williams, M. Davidoff, R. Wang, C. P. Austin, T. P. McDonald, et al., *Biochemical and biophysical research communications* **1999**, *266*, 174–178.

[208] S. Flohr, M. Kurz, E. Kostenis, A. Brkovich, A. Fournier, T. Klabunde, *Journal of medicinal chemistry* **2002**, *45*, 1799–1805.

[209] E. Lescot, J. Sopkova-de Oliveira Santos, C. Dubessy, H. Oulyadi, A. Lesnard, H. Vaudry, R. Bureau, S. Rault, *Journal of chemical information and modeling* **2007**, *47*, 602–612.

[210] P. Grieco, A. Carotenuto, R. Patacchini, C. A. Maggi, E. Novellino, P. Rovero, *Bioorganic & medicinal chemistry* **2002**, *10*, 3731–3739.

[211] A. Carotenuto, P. Grieco, P. Campiglia, E. Novellino, P. Rovero, *Journal of medicinal chemistry* **2004**, *47*, 1652–1661.

[212] E. Haensele, N. Mele, M. Miljak, C. M. Read, D. C. Whitley, L. Banting, C. Delepee, J. Sopkova-de Oliveira Santos, A. Lepailleur, R. Bureau, et al., *Journal of Chemical Information and Modeling* **2017**, *57*, 398–310.

[213] T. Sugo, Y. Murakami, Y. Shimomura, M. Harada, M. Abe, Y. Ishibashi, C. Kitada, N. Miyajima, N. Suzuki, M. Mori, et al., *Biochemical and biophysical research communications* **2003**, *310*, 860–868.

[214] D. Chatenet, C. Dubessy, J. Leprince, C. Boularan, L. Carlier, I. Segalas-Milazzo, L. Guilhaudis, H. Oulyadi, D. Davoust, E. Scalbert, et al., *Peptides* **2004**, *25*, 1819–1830.

[215] H. Vaudry, J.-C. Do Rego, J.-C. Le Mevel, D. Chatenet, H. Tostivint, A. Fournier, M.-C. Tonon, G. Pelletier, J Michael Conlon, J. Leprince, *Annals of the New York Academy of Sciences* **2010**, *1200*, 53–66.

[216] S. M. Foord, T. I. Bonner, R. R. Neubig, E. M. Rosser, J.-P. Pin, A. P. Davenport, M. Spedding, A. J. Harmar, *Pharmacological reviews* **2005**, *57*, 279–288.

[217] R. S. Ames, H. M. Sarau, J. K. Chambers, R. N. Willette, et al., *Nature* **1999**, *401*, 282–286.

[218] A. Stirrat, M. Gallagher, S. A. Douglas, E. H. Ohlstein, C. Berry, A Kirk, M Richardson, M. R. MacLean, *American Journal of Physiology-Heart and Circulatory Physiology* **2001**, *280*, 925–928.

[219] N. A. Elshourbagy, S. A. Douglas, U. Shabon, S. Harrison, G. Duddy, J. L. Sechler, Z. Ao, B. E. Maleeff, D. Naselsky, J. Disa, et al., *British journal of pharmacology* **2002**, *136*, 9–22.

[220] B. Boivin, G. Vaniotis, B. G. Allen, T. E. Hebert, *Journal of Receptors and Signal Transduction* **2008**, *28*, 15–28.

[221] C. Bucharles, P. Bizet, S. Arthaud, A. Arabo, J. Leprince, B. Lefranc, D. Cartier, Y. Anouar, I. Lihrmann, *Journal of Comparative Neurology* **2014**, *522*, 2634–2649.

[222] D Chatenet, M Letourneau, Q. Nguyen, N. Doan, J Dupuis, A Fournier, *British journal of pharmacology* **2013**, *168*, 807–821.